STIC Search Report
Biotech-Chem Library

STIC Database Tracking Number: 109581

TO: Mary K Zeman
Location: cm1/12a17/12d01
Art Unit: 1631
Friday, December 12, 2003

Case Serial Number: 09/940664

From: David Schreiber
Location: Biotech-Chem Library
CM1-6A03
Phone: 308-4292

david.schreiber@uspto.gov

Search Notes

Zeman 09/940,664

=> d his

(FILE 'HOME' ENTERED AT 09:08:14 ON 12 DEC 2003)

FILE 'MEDLINE, HCAPLUS, BIOSIS, EMBASE, SCISEARCH, AGRICOLA' ENTERED AT 09:10:35 ON 12 DEC 2003

```
L1          8752 S NISHIKAWA T?/AU
L2         13096 S MURAKAMI K?/AU
L3           439 S ISOGAI T?/AU
L4         10759 S NAGAI K?/AU
L5         24840 S HAYASHI K?/AU
L6           976 S IRIE R?/AU
L7          1701 S OTSUKI T?/AU
L8         60060 S L1-L7
L9            11 S L8 AND ((3 OR THREE)(3A)FRAME#)
L10         1230 S L8 AND (CDNA OR COMPLEMENT?(A)DNA)
L11          103 S L10 AND FRAME#
L12           13 S L11 AND ALIGN?
L13           28 S L10 AND ALIGN? AND AMINO(A)ACID?
L14           23 S (SOFTWARE? OR ALGORITHM? OR COMPUTER(3A)PROGRAM#) AND ALIGN?(
L15            0 S (SOFTWARE? OR ALGORITHM? OR COMPUTER(3A)PROGRAM#) AND ALIGN?(
L16           63 S (SOFTWARE? OR ALGORITHM? OR COMPUTER(3A)PROGRAM#) AND ALIGN?(
L17           21 S L16 AND (CDNA OR DNA)
L18           29 S (L12 OR L13) NOT PY>2000
L19           73 S L18 OR L14 OR L17
L20           40 DUP REM L19 (33 DUPLICATES REMOVED)
```

=> d ibib abs l20 1-40

L20 ANSWER 1 OF 40 MEDLINE on STN DUPLICATE 1

ACCESSION NUMBER: 2003297178 MEDLINE
DOCUMENT NUMBER: 22708962 PubMed ID: 12824361
TITLE: RevTrans: Multiple **alignment** of coding **DNA** from **aligned amino acid** sequences.
AUTHOR: Wernersson Rasmus; Pedersen Anders Gorm
CORPORATE SOURCE: Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800, Lyngby, Denmark.
SOURCE: NUCLEIC ACIDS RESEARCH, (2003 Jul 1) 31 (13) 3537-9. Journal code: 0411011. ISSN: 1362-4962.
PUB. COUNTRY: England: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 200308
ENTRY DATE: Entered STN: 20030626
Last Updated on STN: 20030819
Entered Medline: 20030818

AB The simple fact that proteins are built from 20 amino acids while DNA only contains four different bases, means that the 'signal-to-noise ratio' in protein sequence alignments is much better than in alignments of DNA. Besides this information-theoretical advantage, protein alignments also benefit from the information that is implicit in empirical substitution matrices such as BLOSUM-62. Taken together with the generally higher rate of synonymous mutations over non-synonymous ones, this means that the phylogenetic signal disappears much more rapidly from DNA sequences than from the encoded proteins. It is therefore preferable to **align** coding **DNA** at the **amino acid** level and it is

Search completed by David Schreiber 308-4292

for this purpose we have constructed the program RevTrans. RevTrans constructs a multiple DNA alignment by: (i) translating the DNA; (ii) aligning the resulting peptide sequences; and (iii) building a multiple DNA alignment by 'reverse translation' of the aligned protein sequences. In the resulting DNA alignment, gaps occur in groups of three corresponding to entire codons, and analogous codon positions are therefore always lined up. These features are useful when constructing multiple DNA alignments for phylogenetic analysis. RevTrans also accepts user-provided protein alignments for greater control of the alignment process. The RevTrans web server is freely available at http://www.cbs.dtu.dk/services/RevTrans/.

L20 ANSWER 2 OF 40 MEDLINE on STN DUPLICATE 2
ACCESSION NUMBER: 2002740609 MEDLINE
DOCUMENT NUMBER: 22391882 PubMed ID: 12503318
TITLE: JavaScript **DNA** translator: **DNA**-aligned protein translations.
AUTHOR: Perry William L 3rd
CORPORATE SOURCE: Lilly Research Laboratories, Lilly Corporate Center, Indianapolis, IN 46285, USA.. bperry@lilly.com
SOURCE: BIOTECHNIQUES, (2002 Dec) 33 (6) 1318-20. Journal code: 8306785. ISSN: 0736-6205.
PUB. COUNTRY: United States
DOCUMENT TYPE: Report; (TECHNICAL REPORT)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 200307
ENTRY DATE: Entered STN: 20021231
Last Updated on STN: 20030716
Entered Medline: 20030715

AB There are many instances in molecular biology when it is necessary to identify ORFs in a **DNA** sequence. While programs exist for displaying protein translations in multiple ORFs in alignment with a **DNA** sequence, they are often expensive, exist as add-ons to **software** that must be purchased, or are only compatible with a particular operating system. JavaScript **DNA** Translator is a shareware application written in JavaScript, a scripting language interpreted by the Netscape Communicator and Internet Explorer Web browsers, which makes it compatible with several different operating systems. While the program uses a familiar Web page interface, it requires no connection to the Internet since calculations are performed on the user's own computer. The program analyzes one or multiple **DNA** sequences and generates translations in up to six reading **frames aligned** to a **DNA** sequence, in addition to displaying translations as separate sequences in FASTA format. ORFs within a reading frame can also be displayed as separate sequences. Flexible formatting options are provided, including the ability to hide ORFs below a minimum size specified by the user. The program is available free of charge at the BioTechniques **Software** Library (www.Biotechniques.com).

L20 ANSWER 3 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN
ACCESSION NUMBER: 2002:438143 HCAPLUS
DOCUMENT NUMBER: 137:305489
TITLE: Organization of the chicken and Xenopus peripherin/rds gene
AUTHOR(S): Li, Chibo; O'Brien, John; Al-Ubaidi, Muayyad R.; Naash, Muna I.
CORPORATE SOURCE: Ophthalmology, Northwestern University, Chicago, IL, 60612, USA

Search completed by David Schreiber 308-4292

SOURCE: New Insights into Retinal Degenerative Diseases, [Proceedings of the International Symposium on Retinal Degeneration], 9th, Durango, CO, United States, Oct. 9-14, 2000 (2001), Meeting Date 2000, 269-277. Editor(s): Anderson, Robert E.; LaVail, Matthew M.; Hollyfield, Joe G. Kluwer Academic/Plenum Publishers: New York, N. Y.
CODEN: 69CSG5; ISBN: 0-306-46679-1
DOCUMENT TYPE: Conference
LANGUAGE: English

AB The exon-intron organization of peripherin/rds from the chicken and the Xenopus was determined Sequence data obtained by direct sequencing were analyzed and edited with the PC/GENE **software**, which was the same program used to generate a comparison of compiled **DNA** sequences and generate multiple **alignments** of predicted **amino acid** sequences from different species available in the GenBank database. Two homologs of peripherin/rds were identified in chicken photoreceptors and three in Xenopus. CRDS1 and XRDS38 are the orthologs of mammalian peripherin/rds while CRDS2, XRDS35 and 36 are more distant relatives and called rds-like proteins.

REFERENCE COUNT: 21 THERE ARE 21 CITED REFERENCES AVAILABLE FOR THIS RECORD. ALL CITATIONS AVAILABLE IN THE RE FORMAT

L20 ANSWER 4 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN DUPLICATE 3
ACCESSION NUMBER: 2001:224034 HCAPLUS
DOCUMENT NUMBER: 136:15747
TITLE: Pro-Frame: similarity-based gene recognition in eukaryotic **DNA** sequences with errors
AUTHOR(S): Mironov, Andrey A.; Novichkov, Pavel S.; Gelfand, Mikhail S.
CORPORATE SOURCE: State Scientific Center for Biotechnology NIIGenetika, Moscow, 113545, Russia
SOURCE: Bioinformatics (2001), 17(1), 13-15
CODEN: BOINFP; ISSN: 1367-4803
PUBLISHER: Oxford University Press
DOCUMENT TYPE: Journal
LANGUAGE: English

AB Performance of existing **algorithms** for similarity-based gene recognition in eukaryotes drops when the genomic **DNA** has been sequenced with errors. A modification of the spliced alignment **algorithm** allows for gene recognition in sequences with errors, in particular frameshifts. It tolerates up to 5% of sequencing errors without considerable drop of prediction reliability when a sufficiently close homologous protein is available (normalized evolutionary distance similarity score 50% or higher).

REFERENCE COUNT: 15 THERE ARE 15 CITED REFERENCES AVAILABLE FOR THIS RECORD. ALL CITATIONS AVAILABLE IN THE RE FORMAT

L20 ANSWER 5 OF 40 BIOSIS COPYRIGHT 2003 BIOLOGICAL ABSTRACTS INC. on STN
ACCESSION NUMBER: 2000:417552 BIOSIS
DOCUMENT NUMBER: PREV200000417552
TITLE: RecA realigns suboptimally paired frames of **DNA** repeats through a process that requires ATP hydrolysis.
AUTHOR(S): Sen, Subhojit; Karthikeyan, G.; Rao, Basuthkar J. [Reprint author]
CORPORATE SOURCE: Department of Biological Sciences, Tata Institute of Fundamental Research, Colaba, Bombay, 400005, India
SOURCE: Biochemistry, (August 22, 2000) Vol. 39, No. 33, pp. 10196-10206. print.

Search completed by David Schreiber 308-4292

CODEN: BICHAW. ISSN: 0006-2960.
DOCUMENT TYPE: Article
LANGUAGE: English
ENTRY DATE: Entered STN: 4 Oct 2000
Last Updated on STN: 8 Jan 2002

AB Microsatellite repeats such as mono-, di-, and trinucleotides are highly abundant and viable targets for homologous recombination in the genome. However, if recombination ensues in such repetitive regions, they are intrinsically prone to frame misalignments during pairing and might eventually give rise to genetic instabilities. Suboptimally paired frames lead to an abrogation of branch migration at the junctions of mixed sequences and repeats, due to a heterologous register. If so, can recombination machinery rectify such misalignments in order to avoid subsequent arrest in branch migration? We analyzed Escherichia coli RecA, the universal prototype of a recombinase, for its pairing abilities across repeats. We used a complementary pairing assay to test whether RecA can mediate realignments of stochastically paired suboptimal **frames** to a maximally **aligned** register. Here, we demonstrate that RecA-single stranded **DNA** filament indeed facilitates such a realignment, probably by sliding the paired strands across mono- and di- as well as trinucleotide repeats. These realignments apparently have no net directional bias. Such a putative "motor" function of RecA seems to be ATP hydrolysis-dependent.

L20 ANSWER 6 OF 40 MEDLINE on STN DUPLICATE 4
ACCESSION NUMBER: 2001211857 MEDLINE
DOCUMENT NUMBER: 21111508 PubMed ID: 11159307
TITLE: Prediction whether a human **cDNA** sequence contains initiation codon by combining statistical information and similarity with protein sequences.
COMMENT: Erratum in: Bioinformatics 2001 Mar;17(3):290
AUTHOR: **Nishikawa T**; Ota T; **Isogai T**
CORPORATE SOURCE: Helix Research Institute, Chiba, Japan.. nisikawa@crl.hitachi.co.jp
SOURCE: BIOINFORMATICS, (2000 Nov) 16 (11) 960-7.
Journal code: 9808944. ISSN: 1367-4803.
PUB. COUNTRY: England: United Kingdom
DOCUMENT TYPE: (EVALUATION STUDIES)
Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 200104
ENTRY DATE: Entered STN: 20010425
Last Updated on STN: 20010723
Entered Medline: 20010419

AB MOTIVATION: In the previous works, we developed ATGpr, a computer program for predicting the fullness of a **cDNA**, i.e. whether it contains an initiation codon or not. Statistical information of short nucleotide fragments was fully exploited in the prediction algorithm. However, sequence similarities to known proteins, which are becoming increasingly available due to recent rapid growth of protein database, were not used in the prediction. In this work, we present a new prediction algorithm based on both statistical and similarity information, which provides better performance in sensitivity and specificity. RESULTS: We evaluated the accuracy of ATGpr for predicting fullness of **cDNA** sequences from human clustered ESTs of UniGene, and we obtained specificity, sensitivity, and correlation coefficient of this prediction. Specificity and sensitivity crossed at 46% over the ATGpr score threshold of 0.33 and the maximum correlation coefficient of 0.34 was obtained at this threshold.

Search completed by David Schreiber 308-4292

Without ATGpr we found it effective to use **alignments** with known proteins for predicting the fullness of **cDNA** sequences. That is, specificity increased monotonously as similarity (identity of the **alignments**) increased. Specificity was achieved greater than 80% if identity was greater than 40%. For more effective prediction of fullness of **cDNA** sequences we combined the similarity (identity of query sequence) with known proteins and ATGpr score. As a result, specificity became greater than 80% if identity was greater than 20%. AVAILABILITY: The prediction program, called ATGpr_ sim, is available at http://www.hri.co.jp/atgpr/ATGpr_sim.html CONTACT: nisikawa@crl.hitachi.co.jp

L20 ANSWER 7 OF 40 MEDLINE on STN DUPLICATE 5
ACCESSION NUMBER: 2000476138 MEDLINE
DOCUMENT NUMBER: 20477694 PubMed ID: 11026670
TITLE: **cDNA** cloning and sequencing of phospholipase A2 from the pyloric ceca of the starfish Asterina pectinifera.
AUTHOR: Kishimura H; Ojima T; **Hayashi K**; Nishita K
CORPORATE SOURCE: Department of Marine Bioresources Chemistry, Faculty of Fisheries, Hokkaido University, Hakodate, Japan.. kishi@fish.hokudai.ac.jp
SOURCE: COMPARATIVE BIOCHEMISTRY AND PHYSIOLOGY. PART B, BIOCHEMISTRY AND MOLECULAR BIOLOGY, (2000 Aug) 126 (4) 579-86.
Journal code: 9516061. ISSN: 1096-4959.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AB022278; GENBANK-AB032266; GENBANK-AB032267
ENTRY MONTH: 200101
ENTRY DATE: Entered STN: 20010322
Last Updated on STN: 20010322
Entered Medline: 20010125

AB Three **cDNA** from the pyloric ceca of the starfish Asterina pectinifera, (namely, **cDNA** 1, 2, and 3), encoding phospholipase A2 (PLA2), were isolated and sequenced. These cDNAs were composed of 415 bp with an open reading **frame** of 414 bp at nucleotide positions 1-414, which encodes 138 **amino acids** including N-terminal Met derived from the PCR primer. The **amino acid** sequence deduced from the **cDNA** 1 was completely consistent with the sequence determined with the starfish PLA2 protein, while those deduced from **cDNA** 2 and **cDNA** 3 differed at one and twelve **amino acid** residual positions, respectively, from the sequence of the PLA2 protein, suggesting the presence of multiple forms in the starfish PLA2. All of the sequences deduced from **cDNA** 1, 2, and 3 required two **amino acid** deletions in pancreatic loop region, and sixteen insertions and three deletions in beta-wing region when **aligned** with the sequence of mammalian pancreatic PLA2. In phylogenetic tree, the starfish PLA2 should be classified into an independent group, but hardly to the established groups IA and IB. The characteristic structure in the pancreatic loop and beta-wing regions may account for the specific properties of the starfish PLA2, e.g. the higher activity and characteristic substrate specificity compared with commercially available PLA2 from porcine pancreas.

L20 ANSWER 8 OF 40 SCISEARCH COPYRIGHT 2003 THOMSON ISI on STN
ACCESSION NUMBER: 2000:626638 SCISEARCH

Search completed by David Schreiber 308-4292

Zeman 09/940,664

THE GENUINE ARTICLE: 342TY
TITLE: **cDNA** cloning and sequencing of phospholipase A(2) from the pyloric ceca of the starfish Asterina pectinifera
AUTHOR: Kishimura H (Reprint); Ojima T; **Hayashi K**; Nishita K
CORPORATE SOURCE: HOKKAIDO UNIV, FAC FISHERIES, DEPT MARINE BIORESOURCES CHEM, HAKODATE, HOKKAIDO 041861, JAPAN (Reprint)
COUNTRY OF AUTHOR: JAPAN
SOURCE: COMPARATIVE BIOCHEMISTRY AND PHYSIOLOGY B-BIOCHEMISTRY & MOLECULAR BIOLOGY, (AUG 2000) Vol. 126, No. 4, pp. 579-586

Publisher: PERGAMON-ELSEVIER SCIENCE LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, ENGLAND.
ISSN: 0305-0491.
DOCUMENT TYPE: Article; Journal
FILE SEGMENT: LIFE
LANGUAGE: English
REFERENCE COUNT: 40
ABSTRACT IS AVAILABLE IN THE ALL AND IALL FORMATS

AB Three **cDNA** from the pyloric ceca of the starfish Asterina pectinifera, (namely, **cDNA** 1, 2, and 3), encoding phospholipase A(2) (PLA(2)), were isolated and sequenced. These cDNAs were composed of 415 bp with an open reading **frame** of 414 bp at nucleotide positions 1-414, which encodes 138 **amino acids** including N-terminal Met derived from the PCR primer. The **amino acid** sequence deduced from the **cDNA** 1 was completely consistent with the sequence determined with the starfish PLA(2) protein, while those deduced from **cDNA** 2 and **cDNA** 3 differed at one and twelve **amino acid** residual positions, respectively, from the sequence of the PLA(2) protein, suggesting the presence of multiple forms in the starfish PLA(2). All of the sequences deduced from **cDNA** 1, 2, and 3 required two **amino acid** deletions in pancreatic loop region, and sixteen insertions and three deletions in beta-wing region when **aligned** with the sequence of mammalian pancreatic PLA(2). In phylogenetic tree, the starfish PLA(2) should be classified into an independent group, but hardly to the established groups IA and IB. The characteristic structure in the pancreatic loop and beta-wing regions may account for the specific properties of the starfish PLA(2), e.g. the higher activity and characteristic substrate specificity compared with commercially available PLA(2) from porcine pancreas. (C) 2000 Elsevier Science Inc. All rights reserved.

L20 ANSWER 9 OF 40 BIOSIS COPYRIGHT 2003 BIOLOGICAL ABSTRACTS INC. on STN
ACCESSION NUMBER: 2000:216292 BIOSIS
DOCUMENT NUMBER: PREV200000216292
TITLE: Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" **cDNA** libraries.
AUTHOR(S): Suzuki, Yutaka [Reprint author]; Ishihara, Daisuke; Sasaki, Masahide; Nakagawa, Haruhito; Hata, Hiroko; Tsunoda, Takeshi; Watanabe, Manabu; Komatsu, Takami; Ota, Toshio; **Isogai, Takao**; Suyama, Akira; Sugano, Sumio
CORPORATE SOURCE: Department of Virology, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
SOURCE: Genomics, (March 15, 2000) Vol. 64, No. 3, pp. 286-297. print.
CODEN: GNMCEP. ISSN: 0888-7543.

Search completed by David Schreiber 308-4292

DOCUMENT TYPE: Article
LANGUAGE: English
ENTRY DATE: Entered STN: 31 May 2000
Last Updated on STN: 5 Jan 2002

AB We constructed 34 types of human "full-length enriched" and "5'-end enriched" **cDNA** libraries based on the "Oligo-Capping" method. We randomly picked and sequenced 10,000 clones from these libraries. BLAST analysis showed that about 50% of the cDNAs were identical to known genes. Among them, we selected 954 species of **cDNA** that should represent the entire sequence from the mRNA start sites. Compared with previously reported sequences, they were on average 45 bp longer in the 5'-end. Using these **cDNA** data, we statistically analyzed the sequence features of the 5'UTR. The average length of the 5'UTR was 125 bp, and there was little correlation with the corresponding mRNA length (correlation coefficiency = 0.26). Of the 954 species of 5'UTR, 459 contained no in-**frame** terminator codon, which is against the common belief. Two hundred seventy-eight species contained at least one ATG codon upstream of the initiator ATG codon. We identified 569 upstream ATGs, in total, 63% of which adequately satisfied Kozak's criteria. These findings are contrary to the typical translation initiation model, which states that translation is initiated from the "first" ATG codon.

L20 ANSWER 10 OF 40 MEDLINE on STN
ACCESSION NUMBER: 2001040432 MEDLINE
DOCUMENT NUMBER: 20435304 PubMed ID: 10978530
TITLE: Isolation and characterization of a novel human gene (NESH) which encodes a putative signaling molecule similar to e3B1 protein.
AUTHOR: Miyazaki K; Matsuda S; Ichigotani Y; Takenouchi Y; **Hayashi K**; Fukuda Y; Nimura Y; Hamaguchi M
CORPORATE SOURCE: Department of Molecular Pathogenesis, Nagoya University School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550, Japan.
SOURCE: BIOCHIMICA ET BIOPHYSICA ACTA, (2000 Sep 7) 1493 (1-2) 237-41.
Journal code: 0217513. ISSN: 0006-3002.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AF037886
ENTRY MONTH: 200012
ENTRY DATE: Entered STN: 20010322
Last Updated on STN: 20010322
Entered Medline: 20001207

AB Using a conventional cloning technique, a novel full-length **cDNA** was isolated and sequenced from a human placental **cDNA** library. This **cDNA** consists of 2129 bp and has a predicted open reading **frame** encoding 366 **amino acids**. It possesses a Src homology 3 (SH3) motif, proline-rich region, serine-rich region and no catalytic domain, suggesting that it seems to be a signaling protein most similar to e3B1, an eps8 SH3 binding protein. PCR-based mapping with both a monochromosomal hybrid panel and radiation hybrid cell panels placed the gene to human chromosome 17q21.3 near the marker D17S1795.

L20 ANSWER 11 OF 40 MEDLINE on STN
ACCESSION NUMBER: 2000386590 MEDLINE
DOCUMENT NUMBER: 20359300 PubMed ID: 10899319
TITLE: Cloning and characterization of rat casein kinase 1epsilon.

Search completed by David Schreiber 308-4292

Zeman 09/940,664

AUTHOR: Takano A; Shimizu K; Kani S; Buijs R M; Okada M; **Nagai K**
CORPORATE SOURCE: Division of Protein Metabolism, Institute for Protein Research, Osaka University, 3-2 Yamado-Oka, Suita, Osaka, Japan.. atsuko@protein.osaka-u.ac.jp
SOURCE: FEBS LETTERS, (2000 Jul 14) 477 (1-2) 106-12.
Journal code: 0155157. ISSN: 0014-5793.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 200008
ENTRY DATE: Entered STN: 20000818
Last Updated on STN: 20000818
Entered Medline: 20000810

AB Genes differentially expressed in the subjective day and night in the rat suprachiasmatic nucleus (SCN) were surveyed by differential display. A gene homologous to human casein kinase 1epsilon (CK1epsilon) was isolated, which initially appeared to be expressed in the suprachiasmatic nucleus (SCN) in a circadian manner. We here describe the **cDNA** cloning of the rat CK1epsilon and characterization of the protein products. The rCK1epsilon is predominantly expressed in the brain including the SCN, binds and phosphorylates mPer1, mPer2, and mPer3 in vitro, and translocates mPer1 and mPer3, but not mPer2, to the cell nucleus depending on its kinase activity when coexpressed with these Per proteins in COS-7 cells.

L20 ANSWER 12 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN
ACCESSION NUMBER: 2000:364768 HCAPLUS
DOCUMENT NUMBER: 133:131713
TITLE: **Amino acid** sequence of phospholipase A2 from the pyloric ceca of starfish Asterina pectinifera
AUTHOR(S): Kishimura, Hideki; Ojima, Takao; Tanaka, Hiroyuki; **Hayashi, Kenji**; Nishita, Kiyoyoshi
CORPORATE SOURCE: Department of Marine Bioresources Chemistry, Faculty of Fisheries, Hokkaido University, Hokkaido, 041-8611, Japan
SOURCE: Fisheries Science (2000), 66(1), 104-109
CODEN: FSCIEH; ISSN: 0919-9268
PUBLISHER: Japanese Society of Fisheries Science
DOCUMENT TYPE: Journal
LANGUAGE: English

AB The complete **amino acid** sequence of phospholipase A2 (PLA2) from the pyloric ceca of the starfish Asterina pectinifera was determined by automated Edman degradation The A. pectinifera PLA2 (APLA2) consists of 137 **amino acids** with an unblocked N-terminus and its mol. weight is calculated to be 15, 300.1. The enzyme contains 14 cysteine (Cys) residues at the corresponding positions of the same residues which have been shown to be involved in intramol. disulfide bonds in mammalian pancreatic PLA2. The region involving an active site and a Ca2+-binding loop shows fairly high sequence homol. (75%) between the APLA2 and porcine pancreatic PLA2. The APLA2 conserved the **amino acid** sequence of the loop portion of the porcine pancreatic PLA2 except for the deletion of two **amino acids**. These features indicate that the APLA2 can be classified into the group 1 type PLA2. In contrast, the homol. between the APLA2 and porcine pancreatic PLA2 was calculated to be 47% in the whole region. Further, the insertion of sixteen residues and

Search completed by David Schreiber 308-4292

the deletion of three residues were required in the sequence of the APLA2 to **align** the corresponding region to the β-wing of porcine pancreatic PLA2. These differences in **amino acid** sequence of the APLA2 may account for its specific properties such as the higher activity and the characteristic substrate specificity.

REFERENCE COUNT: 25 THERE ARE 25 CITED REFERENCES AVAILABLE FOR THIS RECORD. ALL CITATIONS AVAILABLE IN THE RE FORMAT

L20 ANSWER 13 OF 40 BIOSIS COPYRIGHT 2003 BIOLOGICAL ABSTRACTS INC. on STN
ACCESSION NUMBER: 2000:161742 BIOSIS
DOCUMENT NUMBER: PREV200000161742
TITLE: Analysis of messages expressed by Echinostoma paraensei miracidia and sporocysts, obtained by random EST sequencing.
AUTHOR(S): Adema, Coen M. [Reprint author]; Leonard, Pascale M. [Reprint author]; DeJong, Randall J. [Reprint author]; Day, Heather L. [Reprint author]; Edwards, David J. [Reprint author]; Burgett, Georgiana [Reprint author]; Hertel, Lynn A. [Reprint author]; Loker, Eric S. [Reprint author]
CORPORATE SOURCE: Department of Biology, University of New Mexico, Albuquerque, NM, 87131, USA
SOURCE: Journal of Parasitology, (Feb., 2000) Vol. 86, No. 1, pp. 60-65. print.
CODEN: JOPAA2. ISSN: 0022-3395.
DOCUMENT TYPE: Article
LANGUAGE: English
ENTRY DATE: Entered STN: 26 Apr 2000
Last Updated on STN: 4 Jan 2002

AB A lambdaZAP Express **cDNA** library was constructed with mRNA obtained from immature miracidia within eggs, hatched miracidia, and sporocysts of Echinostoma paraensei. This **cDNA** library was amplified and 213 expressed sequence tag (EST) sequences (averaging 466 nucleotides in length) were obtained. The mean percentage of unresolved bases within the EST sequences was 0.4%, ranging from 0 to 4.6%. The 213 ESTs represent 151 unique messages. BLAST (version 2.0.8) analysis disclosed that 64 unique E. paraensei messages (42.4%) had significant similarities (BLAST scoreltoreqe-5), at deduced amino acid or nucleotide levels, with known sequences in the nonredundant GenBank databases or the dbEST database (NCBI). The remainder, 57.6% of the unique EST-encoded messages, scored nonsignificant hits. Most of the E. paraensei messages that could be assigned a cellular role based on sequence similarities were involved in gene/protein expression. Several ESTs scored highest similarities with sequences obtained from trematode species. A total of 22,560 nucleotides present in open reading **frames** from ESTs that **aligned** with known sequences was used to determine codon usage for E. paraensei. Analysis of a subset of eight ESTs that contained full-length open reading frames did not reveal a bias in codon usage. Also, EST sequences were found to contain 3' untranslated regions with an average length of 69.9 +- 88.4 nucleotides (n = 46). The EST sequences were submitted to GenBank/dbEST, adding to the 51 available Echinostoma-derived sequences, to provide reference information for both phylogenetic analysis and study of general trematode biology.

L20 ANSWER 14 OF 40 MEDLINE on STN
ACCESSION NUMBER: 2001646597 MEDLINE
DOCUMENT NUMBER: 21557029 PubMed ID: 11700583
TITLE: An integrated analysis and database system for full-length **cDNA**.
AUTHOR: **Nishikawa T; Murakami K**; Harada N; Ota

Search completed by David Schreiber 308-4292

T; Sugiyama T; **Nagai K; Irie R;** Matui H; Suwa M; **Isogai T**

CORPORATE SOURCE: Biosystems Research Department, Central Research Laboratory, Hitachi, Ltd. 1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan.. nishikawa@crl.hitachi.co.jp

SOURCE: GENOME INFORMATICS SERIES, (2000) 11 12-23.
Journal code: 9717234. ISSN: 0919-9454.
PUB. COUNTRY: Japan
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 200112
ENTRY DATE: Entered STN: 20011112
Last Updated on STN: 20020124
Entered Medline: 20011231

AB Annotation and database system of full-length **cDNA** sequences was developed. As the components of the system, ORF annotation system, functional annotation system based on database search results, mapping annotation system, and integrated retrieval and display system were developed. In the ORF annotation system integrated analyses using conventional tools are performed and useful retrieval interface using motif list are introduced. In the functional annotation system based on database search results, a new method that characterizes a given unknown **cDNA** was developed by using a profile of similarity level over words appearing in sequence database entries. In the mapping annotation system, we linked by similarity searches full-length **cDNA** sequences with database DNA sequences that are already mapped on chromosomes. By using these links, full-length cDNAs can be retrieved by the retrieval condition of physical mapping information. Genetic disease information mapped on the physical mapping site can also be displayed by this system. Furthermore, we constructed an integrated database system for these analyzed data, and thus enabled annotation and selection of full-length cDNAs from points of both gene function and mapping information.

L20 ANSWER 15 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN

ACCESSION NUMBER: 2000:193877 HCAPLUS
DOCUMENT NUMBER: 133:172644
TITLE: FramePlus: aligning **DNA** to protein sequences
AUTHOR(S): Halperin, Eran; Faigler, Simchon; Gill-More, Raveh
CORPORATE SOURCE: Compugen Ltd., Tel Aviv-Jaffa, 69512, Israel
SOURCE: Bioinformatics (1999), 15(11), 867-873
CODEN: BOINFP; ISSN: 1367-4803
PUBLISHER: Oxford University Press
DOCUMENT TYPE: Journal
LANGUAGE: English

Nov 1999

AB Motivation: Automated annotation of Expressed Sequence Tags (ESTs) is becoming increasingly important as EST databases continue to grow rapidly. A common approach to annotation is to align the gene fragments against well-documented databases of protein sequences. The sensitivity of the alignment **algorithm** is key to the success of such methods. Results: This paper introduces a new **algorithm**, **Frame**-Plus, for **DNA**-protein sequence **alignment**. The SCOP database was used to develop a general framework for testing the sensitivity of such alignment **algorithms** when searching large databases. Using this framework, the performance of FramePlus was found to be somewhat better than other **algorithms** in the presence of moderate and high rates of frameshift errors, and comparable to Translated

Search in the absence of sequencing errors. Availability: The source code for FramePlus and the testing datasets are freely available at ftp.compugen.co.il/pub/research. Contact: raveh@compugen.co.il.

REFERENCE COUNT: 24 THERE ARE 24 CITED REFERENCES AVAILABLE FOR THIS RECORD. ALL CITATIONS AVAILABLE IN THE RE FORMAT

L20 ANSWER 16 OF 40 MEDLINE on STN
ACCESSION NUMBER: 2000001940 MEDLINE
DOCUMENT NUMBER: 20001940 PubMed ID: 10529384
TITLE: Activity and substrate specificity of the murine STK2 Serine/Threonine kinase that is structurally related to the mitotic regulator protein NIMA of Aspergillus nidulans.
AUTHOR: **Hayashi K**; Igarashi H; Ogawa M; Sakaguchi N
CORPORATE SOURCE: Department of Immunology, Kumamoto University School of Medicine, 2-2-1, Honjo, Kumamoto, 860-0811, Japan.
SOURCE: BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS, (1999 Oct 22) 264 (2) 449-56.
Journal code: 0372516. ISSN: 0006-291X.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AJ223071; GENBANK-Y09234
ENTRY MONTH: 199912
ENTRY DATE: Entered STN: 20000113
Last Updated on STN: 20020420
Entered Medline: 19991207

AB We isolated a murine STK2 (mSTK2) **cDNA** that is homologous to murine Nek1 serine/threonine kinase, a family member related to the cell cycle regulator kinase NIMA of Aspergillus nidulans. Structural comparison demonstrated that the kinase domain of mSTK2 is highly similar to NIMA/Nek family but the C-terminal region is not similar to any proteins except for human STK2 (hSTK2). Similarly to Nek1, mSTK2 is expressed ubiquitously among various organs and is upregulated in the testis. The expression and localization of mSTK2 are not associated with the cell cycle progression of mitogen-activated lymphocyte and DNA-transfected fibroblast. The substrate specificity of mSTK2 is similar to NIMA, but the phosphorylation is observed exclusively upon threonine residues rather than serine. The mSTK2 is shown to be a new member of the NIMA/Nek family with similar substrate specificity, which might participate in a different role from NIMA kinase involved in the cell cycle regulation.

L20 ANSWER 17 OF 40 BIOSIS COPYRIGHT 2003 BIOLOGICAL ABSTRACTS INC. on STN
ACCESSION NUMBER: 1999:416131 BIOSIS
DOCUMENT NUMBER: PREV199900416131
TITLE: Cloning and expression of chitin deacetylase gene from a deuteromycete, Colletotrichum lindemuthianum.
AUTHOR(S): Tokuyasu, Ken [Reprint author]; Ohnishi-Kameyama, Mayumi; **Hayashi, Kiyoshi**; Mori, Yutaka
CORPORATE SOURCE: National Food Research Institute, Ministry of Agriculture, Forestry and Fisheries, 2-1-2 Kannondai, Tsukuba, Ibaraki, 305-8642, Japan
SOURCE: Journal of Bioscience and Bioengineering, (April, 1999) Vol. 87, No. 4, pp. 418-423. print.
ISSN: 1389-1723.
DOCUMENT TYPE: Article
LANGUAGE: English

Search completed by David Schreiber 308-4292

ENTRY DATE: Entered STN: 18 Oct 1999
Last Updated on STN: 18 Oct 1999

AB The chitin deacetylase gene was cloned from **cDNA** of Colletotrichum lindemuthianum ATCC 56676, and the open reading **frame** consisted of a possible prepro-sequence of 27 **amino acids** at the N-terminus and a mature chitin deacetylase. The deduced **amino acid** sequence of the mature enzyme revealed 26% identity and 46% similarity with a chitin deacetylase from Mucor rouxii. The molecular mass of the protein estimated from the **amino acid** sequence data was 24.3 kDa, which was in good agreement with the MALDI-TOF MS analysis data of the purified protein (24.17-24.36 kDa). The gene product was overexpressed in Escherichia coli cells as a fusion protein with six histidine residues at its C-terminus. The fusion protein formed inclusion bodies, but chitin deacetylase activity was restored from the inclusion bodies by a simple renaturation step with 8 M urea treatment. The recombinant enzyme was purified by affinity chromatography and gel filtration steps, and had a final specific activity of 4.22 units mg-1 of protein. Trypsin digestion of the recombinant enzyme resulted in 2.1-fold increase in activity, suggesting that the removal of the prepro-domain from the recombinant enzyme resulted in an increase in its activity.

L20 ANSWER 18 OF 40 MEDLINE on STN
ACCESSION NUMBER: 1999453741 MEDLINE
DOCUMENT NUMBER: 99453741 PubMed ID: 10524216
TITLE: Molecular cloning and expression of human neurochondrin-1 and -2.
COMMENT: Erratum in: Biochim Biophys Acta 2000 Feb 29;1490(3):367-8
AUTHOR: Mochizuki R; Ishizuka Y; Yanai K; Koga Y; Fukamizu A; **Murakami K**
CORPORATE SOURCE: Sumitomo Pharmaceuticals Research Center, Sumitomo Pharmaceuticals, Osaka, Japan.
SOURCE: BIOCHIMICA ET BIOPHYSICA ACTA, (1999 Sep 3) 1446 (3) 397-402.
Journal code: 0217513. ISSN: 0006-3002.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AB018739; GENBANK-AB018740
ENTRY MONTH: 199911
ENTRY DATE: Entered STN: 20000111
Last Updated on STN: 20000622
Entered Medline: 19991104

AB Human neurochondrins have been cloned from a brain **cDNA** library. The human neurochondrin-1 and -2 predict leucine-rich (15.8 and 15.9%) proteins of 729 and 712 **amino acid** residues, with molecular weights of 78.9 and 77.2 kDa, respectively. The deduced **amino acid** sequence indicates 98% identity among human, mouse and rat species. Northern analysis indicates that about 4 kb human neurochondrin mRNAs are abundant in the fetal and the adult brain.

L20 ANSWER 19 OF 40 MEDLINE on STN
ACCESSION NUMBER: 2000250584 MEDLINE
DOCUMENT NUMBER: 20250584 PubMed ID: 10791922
TITLE: **cDNA** cloning of the two subunits of phospholipase A2 inhibitor PLIgamma from blood plasma of the Chinese mamushi, Agkistrodon blomhoffii siniticus.
AUTHOR: Okumura K; Inoue S; Ohkura N; Ikeda K; **Hayashi K**

Search completed by David Schreiber 308-4292

CORPORATE SOURCE: Department of Biochemistry, Osaka University of Pharmaceutical Sciences, Takatsuki, Japan.
SOURCE: IUBMB Life, (1999 Jul) 48 (1) 99-104.
Journal code: 100888706. ISSN: 1521-6543.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AB018372; GENBANK-AB018373
ENTRY MONTH: 200005
ENTRY DATE: Entered STN: 20000525
Last Updated on STN: 20000525
Entered Medline: 20000516

AB Three phospholipase A2 (PLA2) inhibitors (PLI) have been purified from the blood plasma of the Chinese mamushi, Agkistrodon blomhoffii siniticus; 1 of these, PLIgamma, contains 2 homologous subunits, PLIgamma-A and PLIgamma-B. The cDNAs encoding these 2 subunits of PLIgamma were isolated from a liver **cDNA** library by using fragments from polymerase chain reaction amplifications as probes and sequenced. The respective nucleotide sequences encoded 19-residue signal sequences, followed by 181-residue proteins. The calculated molecular masses were 20123 and 20150 Da for the PLIgamma-A and PLIgamma-B subunits, respectively; and PLIgamma-A included a N-linked carbohydrate site at Asn-157. The sequences of these subunits contained 2 internal repeats of disulfide-bonding pattern characteristic to those of urokinase-type plasminogen activator receptor and members of the Ly-6 superfamily. A phylogenetic analysis comparing the **amino acid** sequences of PLIgamma-A and PLIgamma-B with those for other snakes revealed that the gene duplication leading to these 2 subunits occurred before the divergence of Viperidae and Elapidae.

L20 ANSWER 20 OF 40 MEDLINE on STN DUPLICATE 6
ACCESSION NUMBER: 1999217169 MEDLINE
DOCUMENT NUMBER: 99217169 PubMed ID: 10201112
TITLE: DNA sequence comparison considering both **amino acid** and nucleotide insertions/deletions because of evolution and experimental error.
AUTHOR: **Irie R**; Hiraoka S; Kasahara N; **Nagai K**
CORPORATE SOURCE: Hitachi Ltd., Central Research Laboratory, Tokyo, Japan.. r-irie@crl.hitachi.co.jp
SOURCE: JOURNAL OF BIOTECHNOLOGY, (1999 Mar 26) 69 (1) 19-26.
Journal code: 8411927. ISSN: 0168-1656.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199906
ENTRY DATE: Entered STN: 19990712
Last Updated on STN: 19990712
Entered Medline: 19990623

AB **Amino acid** similarity often needs to be considered in DNA sequence comparison to elucidate gene functions. We propose a Smith-Waterman-like algorithm which considers **amino acid** similarity and insertions/deletions in sequences at the DNA level and at the protein level in a hybrid manner. The algorithm is applied to **cDNA** sequences of Oryza sativa and those of Arabidopsis thaliana. The results are compared with the results of application of NCBI's tblastx program (which compares the sequences in the BLAST manner after translation). It is shown that the present algorithm is very helpful in

discovering nucleotide insertions/deletions originating from experimental errors as well as **amino acid** insertions/deletions due to evolutionary reasons.

L20 ANSWER 21 OF 40 MEDLINE on STN
ACCESSION NUMBER: 1998344034 MEDLINE
DOCUMENT NUMBER: 98344034 PubMed ID: 9677367
TITLE: A novel phospholipase A2 inhibitor with leucine-rich repeats from the blood plasma of Agkistrodon blomhoffii siniticus. Sequence homologies with human leucine-rich alpha2-glycoprotein.
AUTHOR: Okumura K; Ohkura N; Inoue S; Ikeda K; **Hayashi K**
CORPORATE SOURCE: Department of Biochemistry, Osaka University of Pharmaceutical Sciences, Nasahara, Takatsuki, Osaka 569-1094, Japan.
SOURCE: JOURNAL OF BIOLOGICAL CHEMISTRY, (1998 Jul 31) 273 (31) 19469-75.
Journal code: 2985121R. ISSN: 0021-9258.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-AB007198
ENTRY MONTH: 199809
ENTRY DATE: Entered STN: 19980917
Last Updated on STN: 19980917
Entered Medline: 19980910

AB The phospholipase A2 (PLA2) inhibitor PLIbeta, purified from the blood plasma of Chinese mamushi snake (Agkistrodon blomhoffii siniticus), is a 160-kDa trimer with three 50-kDa subunits; and it inhibits specifically the enzymatic activity of the basic PLA2 from its own venom (Ohkura, N., Okuhara, H., Inoue, S., Ikeda, K., and Hayashi, K. (1997) Biochem. J. 325, 527-531). In the present study, the 50-kDa subunit was found to be glycosylated with N-linked carbohydrate, and enzymatic deglycosylation decreased the molecular mass of the 50-kDa subunit to 39-kDa. One 160-kDa trimer of PLIbeta was found to form a stable complex with three basic PLA2 molecules, indicating that one basic PLA2 molecule would bind stoichiometrically to one subunit of PLIbeta. A **cDNA** encoding PLIbeta was isolated from a Chinese mamushi liver **cDNA** library by use of a probe prepared by a polymerase chain reaction on the basis of the partially determined **amino acid** sequence of the subunit. The **cDNA** contained an open reading **frame** encoding a 23-residue signal sequence followed by a 308-residue protein, which contained the sequences of all the peptides derived by lysyl endopeptidase digestion of the subunit. The molecular mass of the mature protein was calculated to be 34,594 Da, and the deduced **amino acid** sequence contained four potential N-glycosylation sites. The sequence of PLIbeta showed no significant homology with that of the known PLA2 inhibitors. But, interestingly, it exhibited 33% identity with that of human leucine-rich alpha2-glycoprotein, a serum protein of unknown function. The most striking feature of the sequence is that it contained nine leucine-rich repeats (LRRs), each of 24 **amino acid** residues and thus encompassing over two-thirds of the molecule. LRRs in PLIbeta might be responsible for the specific binding to basic PLA2, since LRRs are considered as the motifs involved in protein-protein interactions.

L20 ANSWER 22 OF 40 MEDLINE on STN
ACCESSION NUMBER: 1998382529 MEDLINE

Search completed by David Schreiber 308-4292

Zeman 09/940,664

DOCUMENT NUMBER: 98382529 PubMed ID: 9714835
TITLE: Cloning, functional expression, and chromosomal localization of the human and mouse gp180-carboxypeptidase D-like enzyme.
AUTHOR: Ishikawa T; **Murakami K**; Kido Y; Ohnishi S; Yazaki Y; Harada F; Kuroki K
CORPORATE SOURCE: Third Department of Internal Medicine, Faculty of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113, Japan.. takduck-tky@umin.ac.jp
SOURCE: GENE, (1998 Jul 30) 215 (2) 361-70.
Journal code: 7706761. ISSN: 0378-1119.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-D85390; GENBANK-D85391
ENTRY MONTH: 199810
ENTRY DATE: Entered STN: 19981008
Last Updated on STN: 20000303
Entered Medline: 19981001

AB We previously reported that a host cell glycoprotein, gp180, binds duck hepatitis B virus particles, and is encoded by a member of the carboxypeptidase gene family (Kuroki, K., Eng, F., Ishikawa, T., Turck, C., Harada, F., Ganem, D., 1995. gp180, a host cell glycoprotein that binds duck hepatitis B virus particles, is encoded by a member of the carboxypeptidase gene family. J. Biol. Chemical 270, 15022-15028). After that report, carboxypeptidase D (CPD) was subsequently purified from bovine pituitary and characterized as a novel carboxypeptidase E (CPE)-like enzyme, with many characteristics in common with duck gp180 (Song, L., Fricker, L.D., 1995. Purification and characterization of carboxypeptidase D, a novel carboxypeptidase E-like enzyme, from bovine pituitary. J. Biol. Chemical 270, 25007-25013). CPD is now supposed to play an important role in a secretory pathway. To clarify the function of gp180 further, we have isolated and analyzed human and mouse homologues of duck gp180. **cDNA** clones derived from human HepG2 cells and mouse livers have been isolated on the basis of homology to the duck gp180. The suggested open reading **frames** of the human and mouse **cDNA** encode 1380 and 1377 **amino acid** proteins, respectively and have three carboxypeptidase homologous domains (A, B, and C). Domains A and B have completely conserved the residues known to have the enzymatic activity of carboxypeptidase, but domain C in each **cDNA** does not. Northern blotting revealed a ubiquitous tissue distribution of human gp180 mRNA with several transcript species. Expression of human gp180 **cDNA** in transfected 293T<HSP SP = "0. 25">cells exhibited carboxypeptidase activity upon radiometric assay. The human and mouse homologues of duck gp180 have many characteristics in common with bovine CPD. Fluorescence in-situ hybridization reveals that the gene encoding human gp180 is located in region 17q11.2.

L20 ANSWER 23 OF 40 MEDLINE on STN
ACCESSION NUMBER: 1998036124 MEDLINE
DOCUMENT NUMBER: 98036124 PubMed ID: 9370357
TITLE: PCTAIRE 2, a Cdc2-related serine/threonine kinase, is predominantly expressed in terminally differentiated neurons.
AUTHOR: Hirose T; Tamaru T; Okumura N; **Nagai K**; Okada M
CORPORATE SOURCE: Division of Protein Metabolism, Institute for Protein Research, Osaka University, Suita, Japan.. hirose@protein.osaka-u.ac.jp

Search completed by David Schreiber 308-4292

SOURCE: EUROPEAN JOURNAL OF BIOCHEMISTRY, (1997 Oct 15) 249 (2) 481-8.
Journal code: 0107600. ISSN: 0014-2956.
PUB. COUNTRY: GERMANY: Germany, Federal Republic of
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199712
ENTRY DATE: Entered STN: 19980109
Last Updated on STN: 20020420
Entered Medline: 19971212

AB PCTAIRE are members of a subfamily of Cdc2-related kinases that have been shown to be preferentially expressed in post-mitotic cells. To examine the neural functions of PCTAIRE, rat **cDNA** clones encoding PCTAIRE 1, 2, and 3 were isolated, and their expression patterns in the brain were analyzed. Among the three rat PCTAIREs, only PCTAIRE 2 was found to be specifically expressed in the brain. Furthermore, its expression was transiently increased during brain development, peaking 7-15 days after birth. Within the brain, PCTAIRE 2 was concentrated in the neuronal layers of the hippocampus and olfactory bulb, which mostly consist of post-mitotic neurons. In an immunocytochemical experiment, immunoreactivity for PCTAIRE 2 was detected in the cell bodies and extended neurites of neurons, but not in astrocytes. The PCTAIRE 2 protein was recovered in the particulate fraction and resistant to solubilization with non-ionic detergent, suggesting that PCTAIRE 2 might be present as a component of a large protein complex. An immunoprecipitation assay revealed that the PCTAIRE 2 was associated with Ser/Thr-phosphorylating activity for histone H1, and that its activity depended on association with a regulatory partner that can be released under high-salt conditions. These findings suggest that PCTAIRE 2 is a Ser/Thr kinase that might play a unique role in terminally differentiated neurons.

L20 ANSWER 24 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN
ACCESSION NUMBER: 1996:730546 HCAPLUS
DOCUMENT NUMBER: 126:1778
TITLE: Computer analysis of cloned sequences
AUTHOR(S): Caron, Paul R.
CORPORATE SOURCE: Vertex Pharmaceuticals, Cambridge, MA, USA
SOURCE: Methods in Molecular Biology (Totowa, New Jersey) (1997), 69(cDNA Library Protocols), 247-260
CODEN: MMBIED; ISSN: 1064-3745
PUBLISHER: Humana
DOCUMENT TYPE: Journal; General Review
LANGUAGE: English

AB A review with 9 refs. An overview of the types of analyses that should be performed to extract the most information from a sequence is presented. The Genetics Computer Group package is discussed in relation to data entry/fragment assembly, restriction site mapping, finding open reading **frames**, homol. searches, and multiple **alignment**. In addition, functional domain identification and protocols for submission of completed sequences to databanks is discussed. UNIX and VMS operating systems are compared in relation to GCG package. Internet sequence searching is also described where gateways to search engines are presented.

L20 ANSWER 25 OF 40 MEDLINE on STN DUPLICATE 7
ACCESSION NUMBER: 1998066759 MEDLINE
DOCUMENT NUMBER: 98066759 PubMed ID: 9403055

Search completed by David Schreiber 308-4292

TITLE: Comparison of **DNA** sequences with protein sequences.
AUTHOR: Pearson W R; Wood T; Zhang Z; Miller W
CORPORATE SOURCE: Department of Biochemistry, University of Virginia, Charlottesville 22908, USA.. wrp@virginia.EDU
CONTRACT NUMBER: LM04969 (NLM)
LM05110 (NLM)
SOURCE: GENOMICS, (1997 Nov 15) 46 (1) 24-36.
Journal code: 8800135. ISSN: 0888-7543.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199801
ENTRY DATE: Entered STN: 19980129
Last Updated on STN: 19980129
Entered Medline: 19980113

AB The FASTA package of sequence comparison programs has been expanded to include FASTX and FASTY, which compare a **DNA** sequence to a protein sequence database, translating the **DNA** sequence in three **frames** and **aligning** the translated **DNA** sequence to each sequence in the protein database, allowing gaps and frameshifts. Also new are TFASTX and TFASTY, which compare a protein sequence to a **DNA** sequence database, translating each sequence in the **DNA** database in six **frames** and scoring **alignments** with gaps and frameshifts. FASTX and TFASTX allow only frameshifts between codons, while FASTY and TFASTY allow substitutions or frameshifts within a codon. We examined the performance of FASTX and FASTY using different gap-opening, gap-extension, frameshift, and nucleotide substitution penalties. In general, FASTX and FASTY perform equivalently when query sequences contain 0-10% errors. We also evaluated the statistical estimates reported by FASTX and FASTY. These estimates are quite accurate, except when an out-of-frame translation produces a low-complexity protein sequence. We used FASTX to scan the Mycoplasma genitalium, Haemophilus influenzae, and Methanococcus jannaschii genomes for unidentified or misidentified protein-coding genes. We found at least 9 new protein-coding genes in the three genomes and at least 35 genes with potentially incorrect boundaries.

L20 ANSWER 26 OF 40 MEDLINE on STN DUPLICATE 8
ACCESSION NUMBER: 96313239 MEDLINE
DOCUMENT NUMBER: 96313239 PubMed ID: 8759004
TITLE: PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all **DNA** translation frames.
AUTHOR: Birney E; Thompson J D; Gibson T J
CORPORATE SOURCE: European Molecular Biology Laboratory, Heidelberg, Germany.
SOURCE: NUCLEIC ACIDS RESEARCH, (1996 Jul 15) 24 (14) 2730-9.
Journal code: 0411011. ISSN: 0305-1048.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-L04284; GENBANK-L08961; GENBANK-L24895; GENBANK-L33768; GENBANK-M18953; GENBANK-M33166; GENBANK-M33880; GENBANK-M58587; GENBANK-M61877; GENBANK-M96564; GENBANK-M96565; GENBANK-U00061; GENBANK-U00111; GENBANK-U17431; GENBANK-U22181; GENBANK-X12671; GENBANK-X16316; GENBANK-X51315;

Search completed by David Schreiber 308-4292

GENBANK-X51476; GENBANK-X53090; GENBANK-X54530; GENBANK-X73879; GENBANK-X75329; GENBANK-X78116; GENBANK-X78428; SWISSPROT-P13217; SWISSPROT-P13226; SWISSPROT-P13277; SWISSPROT-P17279; SWISSPROT-P18250; +
ENTRY MONTH: 199609
ENTRY DATE: Entered STN: 19960924
Last Updated on STN: 19960924
Entered Medline: 19960917

AB **DNA** translation frames can be disrupted for several reasons, including: (i) errors in sequence determination; (ii) RNA processing, such as intron removal and guide RNA editing; (iii) less commonly, polymerase frameshifting during transcription or ribosomal frameshifting during translation. Frameshifts frequently confound computational activities involving homologous sequences, such as database searches and inferences on structure, function or phylogeny made from multiple alignments. A dynamic alignment **algorithm** is reported here which compares a protein profile (a residue scoring matrix for one or more **aligned** sequences) against the three translation **frames** of a **DNA** strand, allowing frameshifting. The **algorithm** has been incorporated into a new package, WiseTools, for comparison of biological sequences. A protein profile can be compared against either a **DNA** sequence or a protein sequence. The program PairWise may be used interactively for alignment of any two sequence inputs. SearchWise can perform combinations of searches through **DNA** or protein databases by a protein profile or **DNA** sequence. Routine application of the programs has revealed a set of database entries with frameshifts caused by errors in sequence determination.

L20 ANSWER 27 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN
ACCESSION NUMBER: 1996:539463 HCAPLUS
DOCUMENT NUMBER: 125:213485
TITLE: Parametric and inverse-parametric sequence alignment with XPARAL
AUTHOR(S): Gusfield, D.; Stelling, P.
CORPORATE SOURCE: Comput. Sci. Dep., Univ. California, Davis, CA, 95616, USA
SOURCE: Methods in Enzymology (1996), 266(Computer Methods for Macromolecular Sequence Analysis), 481-494
CODEN: MENZAU; ISSN: 0076-6879
PUBLISHER: Academic
DOCUMENT TYPE: Journal
LANGUAGE: English

AB When **aligning DNA** or **amino acid** sequences using numerical-based optimization, there is often considerable disagreement about how to weight matches, mismatches, insertions and deletions, and gaps. Most alignment methods require the user to specify fixed values for those parameters, and it is widely observed that the quality of the resulting alignment can be greatly affected by the choice of parameter settings. The authors here describe a publicly available, user-friendly interactive **software** package, XPARAL, that solves the parametric alignment problem, emphasizing newer features in XPARAL. The use of XPARAL is illustrated by reexamg. an earlier study on gap wts. in protein secondary structure alignment, and the empirical and theor. efficiency of the program is discussed.

L20 ANSWER 28 OF 40 AGRICOLA Compiled and distributed by the National Agricultural Library of the Department of Agriculture of the United States of America. It contains copyrighted materials. All rights reserved. (2003) on STN

Search completed by David Schreiber 308-4292

Zeman 09/940,664

ACCESSION NUMBER: 95:52964 AGRICOLA
DOCUMENT NUMBER: CAT10701056
TITLE: Computer analysis of sequence data.
AUTHOR(S): Griffin, Annette M.; Griffin, Hugh G.
AVAILABILITY: DNAL (QH506.M45 no.24-25)
LC CONTROL NO.: 93-36758 //r94
SOURCE: c1994 2 v. : ill. ; 23 cm
Publisher: Totowa, N.J. : Humana Press, c1994.
Series: Methods in molecular biology (Clifton, N.J.) ; 24-25.
ISBN: 0896032469 (pt. 1), 0896032760 (pt. 2).
NOTE: Includes bibliographical references and indexes.
pt. 1. Computer analysis of sequence data -- GCG: fragment assembly programs -- GCG: drawing linear restriction maps -- GCG: drawing circular restiction maps -- GCG: displaying restriction sites and possible translations in a DNA sequence -- GCG: assembly of sequences into new sequence constructs -- GCG: comparison of sequences -- GCG: production of multiple sequence alignment -- GCG: database searching -- GCG: pattern recognition -- GCG: translation of DNA sequence -- GCG: analysis of protein sequences -- GCG: the analysis of RNA secondary structure -- GCG: preparing sequence data for publication -- MicroGenie: introduction and restriction enzyme analysis -- MicroGenie: shotgun DNA sequencing -- MicroGenie: translation -- MicroGenie: protein analysis.
(cont) MicroGenie: homology searches -- PC/GENE: sequence entry and assembly -- PC/GENE: restriction enzyme analysis -- PC/GENE: translation and searches for protein coding regions -- PC/GENE: sequence comparisons and homologies -- PC/GENE: database searches -- PC/GENE: searches for functional sites in nucleic acids and proteins -- Using the FASTA program to search protein and DNA sequence databases -- Converting between sequence formats -- Obtaining **software** via INTERNET -- Submission of nucleotide sequence data to EMBL/GenBank/DDBJ -- pt. 2. Computer analysis of sequence data -- Staden: introduction -- Staden: sequence input, editing and sequence library use.
(cont) Staden: managing sequence projects -- Staden: statistical and structural analysis of nucleotide sequences -- Staden: searching for restriction sites -- Staden: translating and listing nucleic acid sequences -- Staden: searching for motifs in nucleic acid sequences -- Staden: using patterns to analyze nucleic acid sequences -- Staden: analyzing sequences to find genes -- Staden: statistical and structural analysis of protein sequences -- Staden: searching for motifs in protein sequences -- Staden: using patterns to analyze protein sequences -- Staden: comparing sequences -- Staden plus -- DNA strider: a Macintosh program for handling protein and nucleic acid sequences -- MacVector: an integrated sequence analysis program for the Macintosh -- MacVector: aligning sequences -- MacVector: sequence comparisons using a matrix method.
(cont) MacVector: restriction enzyme analysis --

Search completed by David Schreiber 308-4292

MacVector: protein analysis -- Profile analysis -- Prediction of RNA secondary structure by energy minimization -- Classification and function prediction of proteins using diagnostic **amino acid** patterns -- CLUSTAL V: multiple **alignment** of **DNA** and protein sequences -- Progressive multiple alignment of protein sequences and the construction of phylogenetic trees -- AMPS package for multiple protein sequence alignment -- TreeAlign -- Using the FASTA program to search protein and DNA sequence databases -- Converting between sequence formats -- Obtaining **software** via INTERNET -- Submission of nucleotide sequence data to EMBL/GenBank/DDBJ.

PUB. COUNTRY: New Jersey; United States
DOCUMENT TYPE: Bibliography; (MONOGRAPH)
FILE SEGMENT: U.S. Imprints not USDA, Experiment or Extension
LANGUAGE: English

L20 ANSWER 29 OF 40 MEDLINE on STN DUPLICATE 9
ACCESSION NUMBER: 95007039 MEDLINE
DOCUMENT NUMBER: 95007039 PubMed ID: 7922687
TITLE: A workbench for large-scale sequence homology analysis.
AUTHOR: Sonnhammer E L; Durbin R
CORPORATE SOURCE: Sanger Centre, Hinxton Hall, Cambridge, UK.
SOURCE: COMPUTER APPLICATIONS IN THE BIOSCIENCES, (1994 Jun) 10 (3) 301-7.
Journal code: 8511758. ISSN: 0266-7061.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199411
ENTRY DATE: Entered STN: 19941222
Last Updated on STN: 19941222
Entered Medline: 19941118

AB When routinely analysing very long stretches of **DNA** sequences produced by genome sequencing projects, detailed analysis of database search results becomes exceedingly time consuming. To reduce the tedious browsing of large quantities of protein similarities, two programs, MSPcrunch and Blixem, were developed, which assist in processing the results from the database search programs in the BLAST suite. MSPcrunch removes biased composition and redundant matches while keeping weak matches that are consistent with a larger gapped alignment. This makes BLAST searching in practice more sensitive and reduces the risk of overlooking distant similarities. Blixem is a multiple sequence alignment viewer for X-windows which makes it significantly easier to scan and evaluate the matches ratified by MSPcrunch. In Blixem, matches to the translated **DNA** query sequence are simultaneously **aligned** in three **frames**. Also, the distribution of matches over the whole **DNA** query is displayed. Examples of usage are drawn from 36 C. elegans cosmid clones totalling 1.2 megabases, to which these tools were applied.

L20 ANSWER 30 OF 40 MEDLINE on STN DUPLICATE 10
ACCESSION NUMBER: 94118255 MEDLINE
DOCUMENT NUMBER: 94118255 PubMed ID: 8289235
TITLE: Sequence alignment and penalty choice. Review of concepts, case studies and implications.

AUTHOR: Vingron M; Waterman M S
CORPORATE SOURCE: Department of Mathematics, University of Southern California, Los Angeles 90089-1113.
CONTRACT NUMBER: GM36230 (NIGMS)
SOURCE: JOURNAL OF MOLECULAR BIOLOGY, (1994 Jan 7) 235 (1) 1-12.
Ref: 24
Journal code: 2985088R. ISSN: 0022-2836.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
General Review; (REVIEW)
(REVIEW, TUTORIAL)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199402
ENTRY DATE: Entered STN: 19940312
Last Updated on STN: 19940312
Entered Medline: 19940224

AB **Alignment algorithms** to compare **DNA** or **amino acid** sequences are widely used tools in molecular biology. The **algorithms** depend on the setting of various parameters, most notably gap penalties. The effect that such parameters have on the resulting alignments is still poorly understood. This paper begins by reviewing two recent advances in **algorithms** and probability that enable us to take a new approach to this question. The first tool we introduce is a newly developed method to delineate efficiently all optimal alignments arising under all choices of parameters. The second tool comprises insights into the statistical behavior of optimal alignment scores. From this we gain a better understanding of the dependence of alignments on parameters in general. We propose novel criteria to detect biologically good alignments and highlight some specific features about the interaction between similarity matrices and gap penalties. To illustrate our analysis we present a detailed study of the comparison of two immunoglobulin sequences.

L20 ANSWER 31 OF 40 MEDLINE on STN
ACCESSION NUMBER: 93327934 MEDLINE
DOCUMENT NUMBER: 93327934 PubMed ID: 8335106
TITLE: Identification of an isoform with an extremely large Cys-rich region of PC6, a Kex2-like processing endoprotease.
AUTHOR: Nakagawa T; **Murakami K**; Nakayama K
CORPORATE SOURCE: Institute of Applied Biochemistry, University of Tsukuba, Ibaraki, Japan.
SOURCE: FEBS LETTERS, (1993 Jul 26) 327 (2) 165-71.
Journal code: 0155157. ISSN: 0014-5793.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199308
ENTRY DATE: Entered STN: 19930903
Last Updated on STN: 20000303
Entered Medline: 19930824

AB In the previous study [1993, J. Biochem. (Tokyo) 113, 132-135] we identified PC6, a member of the Kex2 family of processing endoproteases. In this study, we identified another **cDNA** encoding an isoform of PC6, and designated it as PC6B and redesignated the originally identified PC6 as PC6A. PC6B had a very large Cys-rich region consisting of 22-times repeats of a Cys-rich motif, and a putative transmembrane domain which is

not present in PC6A. A PC6B transcript was found mainly in the intestine, while PC6A transcripts were in various tissues. These results suggest distinct roles of PC6A and PC6B in endoproteolytic processing of precursor proteins.

L20 ANSWER 32 OF 40 MEDLINE on STN
ACCESSION NUMBER: 92292177 MEDLINE
DOCUMENT NUMBER: 92292177 PubMed ID: 1602493
TITLE: Early evolutionary relationships among known life forms inferred from elongation factor EF-2/EF-G sequences: phylogenetic coherence and structure of the archaeal domain.
AUTHOR: Cammarano P; Palm P; Creti R; Ceccarelli E; Sanangelantoni A M; Tiboni O
CORPORATE SOURCE: Istituto Pasteur-Fondazione Cenci Bolognetti, Dipartimento di Biopatologia Umana, Universita di Roma, La Sapienza, Roma, Italy.
SOURCE: JOURNAL OF MOLECULAR EVOLUTION, (1992 May) 34 (5) 396-405. Journal code: 0360051. ISSN: 0022-2844.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals; Space Life Sciences
ENTRY MONTH: 199207
ENTRY DATE: Entered STN: 19920724
Last Updated on STN: 20000303
Entered Medline: 19920715

AB Phylogenies were inferred from both the gene and the protein sequences of the translational elongation factor termed EF-2 (for Archaea and Eukarya) and EF-G (for Bacteria). All treeing methods used (distance-matrix, maximum likelihood, and parsimony), including evolutionary parsimony, support the archaeal tree and disprove the "eocyte tree" (i.e., the polyphyly and paraphyly of the Archaea). Distance-matrix trees derived from both the **amino acid** and the **DNA** sequence **alignments** (first and second codon positions) showed the Archaea to be a monophyletic-holophyletic grouping whose deepest bifurcation divides a Sulfolobus branch from a branch comprising Methanococcus, Halobacterium, and Thermoplasma. Bootstrapped distance-matrix treeing confirmed the monophyly-holophyly of Archaea in 100% of the samples and supported the bifurcation of Archaea into a Sulfolobus branch and a methanogen-halophile branch in 97% of the samples. Similar phylogenies were inferred by maximum likelihood and by maximum (protein and DNA) parsimony. DNA parsimony trees essentially identical to those inferred from first and second codon positions were derived from alternative DNA data sets comprising either the first or the second position of each codon. Bootstrapped DNA parsimony supported the monophyly-holophyly of Archaea in 100% of the bootstrap samples and confirmed the division of Archaea into a Sulfolobus branch and a methanogen-halophile branch in 93% of the bootstrap samples. Distance-matrix and maximum likelihood treeing under the constraint that branch lengths must be consistent with a molecular clock placed the root of the universal tree between the Bacteria and the bifurcation of Archaea and Eukarya. The results support the division of Archaea into the kingdoms Crenarchaeota (corresponding to the Sulfolobus branch and Euryarchaeota). This division was not confirmed by evolutionary parsimony, which identified Halobacterium rather than Sulfolobus as the deepest offspring within the Archaea.

L20 ANSWER 33 OF 40 BIOSIS COPYRIGHT 2003 BIOLOGICAL ABSTRACTS INC. on STN

Search completed by David Schreiber 308-4292

ACCESSION NUMBER: 1992:477825 BIOSIS
DOCUMENT NUMBER: PREV199294109200; BA94:109200
TITLE: LOCAL MULTIPLE ALIGNMENT BY CONSENSUS MATRIX.
AUTHOR(S): ALEXANDROV N N [Reprint author]
CORPORATE SOURCE: CHEM DEP, FAC SCI, KYOTO UNIV, KYOTO 606, JPN
SOURCE: Computer Applications in the Biosciences, (1992) Vol. 8, No. 4, pp. 339-345.
DOCUMENT TYPE: Article
FILE SEGMENT: BA
LANGUAGE: ENGLISH
ENTRY DATE: Entered STN: 27 Oct 1992
Last Updated on STN: 27 Oct 1992

AB A new **algorithm** for aligning several sequences based on the calculation of a consensus matrix and the comparison of all the sequences using this consensus matrix is described. This consensus matrix contains the preference scores of each nucleotide/amino acid and gaps in every position of the alignment. Two modifications of the **algorithm** corresponding to the evolutionary and functional meanings of the alignment were developed. The first one solves the best-fitting problem without any penalty for end gaps and with an internal gap penalty function independent on the gap length. This **algorithm** should be used when comparing evolutionary-related proteins for identifying the most conservative residues. The other modification of the **algorithm** finds the most similar segments in the given sequences. It can be used for finding those parts of the sequences that are responsible for the same biological function. In this case the gap penalty function was chosen to be proportional to the gap length. The results of aligning amino acid sequences of neutral proteases and a compilation of 65 allosteric effectors and substrates of PEP carboxylase are presented.

L20 ANSWER 34 OF 40 MEDLINE on STN
ACCESSION NUMBER: 92275088 MEDLINE
DOCUMENT NUMBER: 92275088 PubMed ID: 1317302
TITLE: Multiple genes for Xenopus activin receptor expressed during early embryogenesis.
AUTHOR: Nishimatsu S; Oda S; **Murakami K**; Ueno N
CORPORATE SOURCE: Institute of Applied Biochemistry, University of Tsukuba, Ibaraki, Japan.
SOURCE: FEBS LETTERS, (1992 May 25) 303 (1) 81-4.
Journal code: 0155157. ISSN: 0014-5793.
PUB. COUNTRY: Netherlands
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199206
ENTRY DATE: Entered STN: 19920710
Last Updated on STN: 19970203
Entered Medline: 19920630

AB Four distinct cDNAs for activin receptor designated as XSTK2, 3, 8 and 9 have been cloned from a Xenopus laevis **cDNA** library. The protein structures deduced from the cDNAs have shown that they all have a putative extracellular ligand-binding domain, a single transmembrane domain and cytoplasmic Ser/Thr kinase domain, except that XSTK2 is extremely similar to the XSTK3 gene but lacks a carboxyl-terminal part of the kinase motif. Northern blot analysis showed that all transcripts are maternally inherited. The levels of transcript for XSTK2, 3 and 8 appeared to fluctuate during early development while those for XSTK9 maintain constant.

Search completed by David Schreiber 308-4292

L20 ANSWER 35 OF 40 MEDLINE on STN
ACCESSION NUMBER: 92011720 MEDLINE
DOCUMENT NUMBER: 92011720 PubMed ID: 1918045
TITLE: Mouse submandibular gland prorenin-converting enzyme is a member of glandular kallikrein family.
AUTHOR: Kim W S; Nakayama K; Nakagawa T; Kawamura Y; Haraguchi K; **Murakami K**
CORPORATE SOURCE: Institute of Applied Biochemistry, University of Tsukuba, Ibaraki, Japan.
SOURCE: JOURNAL OF BIOLOGICAL CHEMISTRY, (1991 Oct 15) 266 (29) 19283-7.
Journal code: 2985121R. ISSN: 0021-9258.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
OTHER SOURCE: GENBANK-M69219; GENBANK-M72403; GENBANK-M72404; GENBANK-M72405; GENBANK-M72406; GENBANK-M72407; GENBANK-M72408; GENBANK-S58334; GENBANK-S58340; GENBANK-S58401; GENBANK-X58628
ENTRY MONTH: 199111
ENTRY DATE: Entered STN: 19920124
Last Updated on STN: 20000303
Entered Medline: 19911114

AB Mouse submandibular gland prorenin-converting enzyme (PRECE) consists of the two polypeptide chains of 17 and 10 kDa and cleaves mouse Ren-2 prorenin at a dibasic site to yield mature renin. Western blot analysis using an antiserum against this enzyme gave rise to multiple bands in mouse submandibular glands, suggesting that PRECE is a member of a protease family. Partial **amino acid** sequence analysis of purified PRECE and cloning and sequence analyses of its **cDNA** indicated that it is identical to the mGK-13 gene product, epidermal growth factor-binding protein type B, which is a member of the glandular kallikrein family and is involved in maturation of epidermal growth factor. Conditioned medium from Chinese hamster ovary cells transfected with an expression plasmid for PRECE had prorenin converting activity. These results indicate that PRECE is involved in the maturation of two bioactive polypeptides expressed in mouse submandibular glands, Ren-2 renin and epidermal growth factor.

L20 ANSWER 36 OF 40 MEDLINE on STN
ACCESSION NUMBER: 92078089 MEDLINE
DOCUMENT NUMBER: 92078089 PubMed ID: 1744039
TITLE: Molecular analysis of Bacillus subtilis ada mutants deficient in the adaptive response to simple alkylating agents.
AUTHOR: Morohoshi F; **Hayashi K**; Munakata N
CORPORATE SOURCE: Radiobiology Division, National Cancer Center Research Institute, Tokyo, Japan.
SOURCE: JOURNAL OF BACTERIOLOGY, (1991 Dec) 173 (24) 7834-40.
Journal code: 2985120R. ISSN: 0021-9193.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 199201
ENTRY DATE: Entered STN: 19920202
Last Updated on STN: 19920202
Entered Medline: 19920115

Search completed by David Schreiber 308-4292

AB Previously, we isolated and characterized six Bacillus subtilis ada mutants that were hypersensitive to methylnitroso compounds and deficient in the adaptive response to alkylation. Cloning of the **DNA complementing** the defects revealed the presence of an ada operon consisting of two tandem and partially overlapping genes, adaA and adaB. The two genes encoded proteins with methylphosphotriester-DNA methyltransferase and O6-methylguanine-DNA methyltransferase activities, respectively. To locate the six mutations, the ada operon was divided into five overlapping regions of about 350 bp. The fragments of each region were amplified by polymerase chain reaction and analyzed by gel electrophoresis to detect single-strand conformation polymorphism. Nucleotide sequences of the fragments exhibiting mobility shifts were determined. Three of the mutants carried sequence alterations in the adaA gene: the adaA1 and adaA2 mutants had a one-base deletion and insertion, respectively, and the adaA5 mutant had a substitution of two consecutive bases causing changes of two **amino acid** residues next to the presumptive alkyl-accepting Cys-85 residue. Three mutants carried sequence alterations in the adaB gene: the adaB3 mutant contained a rearrangement, the adaB6 mutant contained a base substitution causing a change of the presumptive alkyl-accepting Cys-141 to Tyr, and the adaB4 mutant contained a base substitution changing Leu-167 to Pro. The adaB mutants produced ada transcripts upon treatment with low doses of alkylating agents, whereas the adaA mutant did not. We conclude that the AdaA protein functions as the transcriptional activator of this operon, while the AdaB protein specializes in repair of alkylated residues in DNA.

L20 ANSWER 37 OF 40 HCAPLUS COPYRIGHT 2003 ACS on STN DUPLICATE 11
ACCESSION NUMBER: 1989:511586 HCAPLUS
DOCUMENT NUMBER: 111:111586
TITLE: Cloning and sequence analysis of **cDNA** for Irpex lacteus aspartic proteinase
AUTHOR(S): Kobayashi, Hideyuki; Sekibata, Satoshi; Shibuya, Hiroshi; Yoshida, Shigeki; Kusakabe, Isao; **Murakami, Kazuo**
CORPORATE SOURCE: Inst. Appl. Biochem., Univ. Tsukuba, Ibaraki, 305, Japan
SOURCE: Agricultural and Biological Chemistry (1989), 53(7), 1927-33
CODEN: ABCHA6; ISSN: 0002-1369
DOCUMENT TYPE: Journal
LANGUAGE: English
AB To find the primary sequence of I. lacteus aspartic proteinase (ILAP), a **cDNA** library of I. lacteus mRNA in pBR322 was constructed. A clone, which had an insert size of about 1.2 kilobase pairs, was found to contain the coding region of the mature enzyme. The deduced **amino acid** sequence showed that the enzyme consisted of 340 **amino acid** residues with a mol. weight of 35,000. Cysteine and methionine were not found in the enzyme, and 2 putative N-glycosylation sites were indicated. The lack of S-S bridges in the mol. is a striking feature of the enzyme. The **alignment** of the sequence of the enzyme against other aspartic proteinases revealed homol. around the active site aspartic acid residues.

L20 ANSWER 38 OF 40 MEDLINE on STN DUPLICATE 12
ACCESSION NUMBER: 89089252 MEDLINE
DOCUMENT NUMBER: 89089252 PubMed ID: 3208181
TITLE: Multiple DNA and protein sequence alignment on a workstation and a supercomputer.
AUTHOR: Tajima K

Search completed by David Schreiber 308-4292

CORPORATE SOURCE: Biological Informatics Section, International Institute for Advanced Study of Social Information Science, Tokyo, Japan.
SOURCE: COMPUTER APPLICATIONS IN THE BIOSCIENCES, (1988 Nov) 4 (4) 467-71.
Journal code: 8511758. ISSN: 0266-7061.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals; AIDS
ENTRY MONTH: 198902
ENTRY DATE: Entered STN: 19900308
Last Updated on STN: 19970203
Entered Medline: 19890216

AB This paper describes a multiple alignment method using a workstation and supercomputer. The method is based on the alignment of a set of aligned sequences with the new sequence, and uses a recursive procedure of such alignment. The alignment is executed in a reasonable computation time on diverse levels from a workstation to a supercomputer, from the viewpoint of alignment results and computational speed by parallel processing. The application of the **algorithm** is illustrated by several examples of multiple **alignment** of 12 **amino acid** and **DNA** sequences of HIV (human immunodeficiency virus) env genes. Colour graphic programs on a workstation and parallel processing on a supercomputer are discussed.

L20 ANSWER 39 OF 40 MEDLINE on STN
ACCESSION NUMBER: 89144951 MEDLINE
DOCUMENT NUMBER: 89144951 PubMed ID: 3067216
TITLE: Computer-aided detection and alignment of weakly homologous amino acid sequences of RNA replicase beta (MS2 phage) and DNA polymerases (T7 phage and E. coli).
AUTHOR: Ohnishi K
CORPORATE SOURCE: Department of Biology, Faculty of Science, Niigata University, Japan.
SOURCE: NUCLEIC ACIDS SYMPOSIUM SERIES, (1988) (19) 193-7.
Journal code: 8007206. ISSN: 0261-3166.
PUB. COUNTRY: ENGLAND: United Kingdom
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 198903
ENTRY DATE: Entered STN: 19900306
Last Updated on STN: 19980206
Entered Medline: 19890329

AB **Alignment** of the **amino acid** (aa) sequences of T7 phage **DNA** polymerase (DPase), E. coli DNA polymerase I (Pol I) and MS2 phage RNA replicase beta subunit (MS2 Repl) were established by computer-aided methods. The results showed that the entire length (aa's 16-704) of T7 DPase is homologous to Pol I aa's 207-928(C-term) with 21.5% aa identity, and that domains I (aa's-1-311) and II (312-451(C-term] were found to be homologous to each other and to N-terminal region of T7 DPase (aa's 1-250). Thus these enzymes and domains are homologous to one another and must have evolved from a co-ancestral enzyme.

L20 ANSWER 40 OF 40 MEDLINE on STN DUPLICATE 13
ACCESSION NUMBER: 85063753 MEDLINE
DOCUMENT NUMBER: 85063753 PubMed ID: 6594689
TITLE: Internal duplication in human alpha 1 and beta 1

Search completed by David Schreiber 308-4292

interferons.

AUTHOR: Erickson B W; May L T; Sehgal P B

CONTRACT NUMBER: AI 16262 (NIAID)
GM 32622 (NIGMS)
SOURCE: PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, (1984 Nov) 81 (22) 7171-5.
Journal code: 7505876. ISSN: 0027-8424.
PUB. COUNTRY: United States
DOCUMENT TYPE: Journal; Article; (JOURNAL ARTICLE)
LANGUAGE: English
FILE SEGMENT: Priority Journals
ENTRY MONTH: 198501
ENTRY DATE: Entered STN: 19900320
Last Updated on STN: 19970203
Entered Medline: 19850102

AB Metric analysis of the nucleotide sequence of the intron-free human interferon beta 1 (IFN-beta 1) gene by using the Sellers TT **algorithm** revealed that this gene contains two major repeated segments, which span the entire coding region. These repeats are each approximately 300 nucleotides in length and have 45% identical aligned nucleotides (common bases). When these metrically **aligned DNA** repeats were translated into **amino acids**, 9 (19%) of the 47 in-phase amino acid residues were identical (common acids). This internal duplication was also apparent on visual inspection of the amino acid sequence of IFN-beta 1. In addition, metric analysis of the nucleotide sequence of the intron-free IFN-alpha 1 gene showed that this gene also contains two repeats, each approximately 300 nucleotides long, having 47% common bases and 19% common acids. Since the IFN-alpha 1 and -beta 1 genes are known to be related (by the present metric analysis they contain 53% common bases and 45% common acids), a consensus DNA sequence was derived from all four of these repeats. Manual alignment of the separate metric alignments corresponding to the two halves of the IFN-alpha 1 and -beta 1 genes provided a composite alignment with 58% of the alignment positions having the same nucleotide in at least three of the four repeats. When this composite nucleotide alignment was translated to define a composite alignment of the four protein segments, 10 (31%) of the 32 in-phase amino acid residues contained the same amino acid in at least three of the four segments. These sequences relationships provide insight into the origin of the IFN-alpha 1 and -beta 1 genes and furnish an additional basis for comparing them with other related genes.

Products
- Drug Target Discovery Platform
- Genecarta™
- GeneGuide™
- Probe Design and Analysis
- OligoLibrary™
- Proteomics
- Novel Gene Portfolio
- Database Search Products
  - Hardware
  - Software

Products — Software

This part describes GenCore programs that search sequence databases. FASTA and BLAST families of applications perform heuristic homology searches and provide statistical estimations of the results. Search applications incorporated in the OneModel paradigm allow for more sensitive rigorous database searches, such as Smith-Waterman (including Profilesearch) and FrameSearch.

The OneModel Application

Description

The OneModel application is Compugen's scheme for easily describing dynamic-programming algorithms (Hidden Markov Models) using a model-definition file. OneModel is a generic application that performs the calculation of score and alignment in accordance with the algorithm defined in the model file. In future versions, the OneModel scheme would enable you to program a new dynamic-programming search application without writing any C or other programming language code. For further details about OneModel see **OneModel Tech Info**.

FrameSearch Models

The following table describes the FrameSearch models supplied by Compugen and the query and database types that you can use for the search with the specified model:

frame_n2p	Three-state model allowing for frameshifts that is based on the FrameSearch algorithm developed by GCG[9].	Compares a DNA query to a protein database.
frame_p2n	Three-state model allowing for frameshifts that is based on the FrameSearch algorithm developed by GCG. Includes also **proframesearch.**	Compares a protein query sequence/profile to a DNA database.

Table 11: Models supplied by Compugen Ltd. with the OneModel application.

FrameSearch Models Acceleration

frame_n2p, frame_p2n (includes also proframesearch)

Implementation	BioXL/G	Bioccelerator	Software-only
frame_n2p	+	+	+
frame_p2n, proframesearch	+	+	+

frame_n2p and frame_p2n models are based on the FrameSearch algorithm developed by GCG[9]. FrameSearch aligns the protein sequence to the codons of the nucleic sequence in all possible reading frames, allowing for alignments that include reading frame shift errors in the nucleic sequence.

You can use **frame_n2p** to compare a DNA query to a protein database.
frame_p2ncompares a protein sequence/profile to a DNA database.

Usage for searches with nucleic acid queries:

```
om -model=frame_n2p [-q=]query [-db=]database [options]
```

or

```
frame_n2p [-q=]query [-db=]database [options]
```

Usage for searches with protein queries:

```
om -model=frame_p2n [-q=]query [-db=]database [options]
```

or

```
frame_p2n [-q=]query [-db=]database [options]
```

or

```
proframesearch [-q=]query [-db=]database [options]
```

9 Edelman, et al., "A rigorous program for searching protein databases with nucleic acid queries," poster, Genome Sequence and Analysis Conference, Hilton Head, 1995.

FRAMESEARCH*(+)

Go back to top

- -YPAN=30.0
- -PORtrait

FUNCTION

FrameSearch searches a group of protein sequences for similarity to one or more nucleotide query sequences, or searches a group of nucleotide sequences for similarity to one or more protein query sequences. For each sequence comparison, the program finds an optimal alignment between the protein sequence and all possible codons on each strand of the nucleotide sequence. Optimal alignments may include reading frame shifts.

DESCRIPTION

FrameSearch searches a group of protein sequences for similarity to one or more nucleotide query sequences, or searches a group of nucleotide sequences for similarity to one or more protein query sequences. For each sequence comparison, the program creates the optimal local alignment of the best region of similarity between the protein sequence and all possible codons on each strand of the nucleotide sequence. Because FrameSearch can match the protein to codons in different reading frames of the nucleotide sequence as part of the same alignment, it can identify sequence similarity even when the nucleotide sequence contains reading frame shifts.

In standard sequence alignment programs, you routinely specify gap creation and extension penalties. In addition to these penalties, FrameSearch also allows you to specify a separate frameshift penalty for the creation of gaps that result in reading frame shifts in the nucleotide sequence. (See the ALGORITHM topic for a more detailed explanation of how gaps are penalized.)

By default, the search proceeds as a local alignment between the query sequence and each sequence in the search set. Optionally, you can search using a global alignment procedure where FrameSearch inserts gaps to optimize the alignment between the entire nucleotide sequence and the entire protein sequence.

The search output contains an ordered list of the sequences in the search set that have the highest comparison scores when aligned to the query sequence. The actual alignments for these top-scoring matches are displayed after the list.

You can specify multiple query sequences (such as a list file or a sequence specification using an asterisk (*) wildcard) as input to FrameSearch. The program compares each query sequence separately to the sequences specified in the search set, and it writes a separate output file for each query search. If you use a list file as your query, you can add begin and end sequence attributes to specify the range for each query sequence. For more information about list files, see "Using List Files (formerly Files of Sequence Names) " in Chapter 2, Using Sequences in the User's Guide.

EXAMPLE

Here is a session using FrameSearch to find sequences in SWISS-PROT with similarities to the translation product of the cDNA sequence EST:Atts0012.

```
% FrameSearch -PLOt

 FRAMESEARCH with what query sequence(s) ? EST:Atts0012

                  Begin (* 1 *) ?
                End (*   286 *) ?
```

```
Search for query in what sequence(s) (* SwissProt:* *) ?

What is the gap creation penalty (* 12.00 *) ?

What is the gap extension penalty (* 4.00 *) ?

What is the frameshift penalty (* 0.00 *) ?

This program can plot the distribution of alignment search scores graphically.
Do you want to:

    A) Plot to a FIGURE file called "framesearch.figure"
    B) Plot graphics on LaserWriter attached to /dev/tty10

Please choose one (* A *):

What should I call the output file (* atts0012.framesearch *) ?

         1 Sequences          924 aa searched    SW:104kthepa
       101 Sequences       36,727 aa searched    SW:1a38human

       ////////////////////////////////////////////////////


    43,301 Sequences  15,271,754 aa searched    SW:Z123HUMAN
    43,401 Sequences  15,311,594 aa searched    SW:ZN15HUMAN

Aligning.....................................

FIGURE instructions are now being written into framesearch.figure.


CPU time used:
       Search time:  2:28: 6.2
  Post-search time:  0: 0: 6.4
    Total CPU time:  2:28:12.6

Output File: atts0012.framesearch

%
```

OUTPUT

Here is some of the output file:

```
  FRAMESEARCH of: GB_EST:ATTS0012  check: 2422   from: 1  to: 286

LOCUS       ATTS0012      286 bp    RNA             EST       31-OCT-1992
DEFINITION  A. thaliana transcribed sequence; clone TAT1B11, 5' end; similar
            to GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE.
ACCESSION   Z17438
KEYWORDS    expressed sequence tag; partial cDNA sequence.
SOURCE      thale cress. . . .


 TO: SwissProt:*  Sequences: 43,479  Total-length: 15,335,248
      March 8, 1995 15:40

 Scoring matrix: GenRunData:framepep.cmp
 Translation table: GenRunData:translate.txt

  Gap creation penalty: 12.000
 Gap extension penalty:  4.000
```

```
    Frameshift penalty:  0.000

The best scores are:                                                    ..

sw:g3pc_arath  P25858 arabidopsis thaliana (mouse-ear cress). glyce...  343.0
sw:g3pc_sinal  P04796 sinapis alba (white mustard). glyceraldehyde ...  331.0
sw:g3pc_ranac  P26521 ranunculus acer (common buttercup). glycerald...  313.0

//////////////////////////////////////////////////////////////////////////////

sw:g3p_schco  P32638 schizophyllum commune (bracket fungus). glycer...  227.0
sw:g3p_pig  P00355 sus scrofa (pig). glyceraldehyde 3-phosphate deh...  227.0
sw:g3p_klula  P17819 kluyveromyces lactis (yeast). glyceraldehyde 3...  227.0

atts0012
g3pc_arath

            Quality:  343.0              Length:    240
              Ratio:  4.397                Gaps:      2
 Percent Similarity: 100.000   Percent Identity: 97.436

                .         .         .         .         .
      3 GAAATCAAGAAGGCCATCAAGGAGGAATCTGAAGGCAAAATGAAGGGAAT 52
        |||||||||||||||||||||||||||||||||||||:::||||||||
    261 GluIleLysLysAlaIleLysGluGluSerGluGlyLysLeuLysGlyIl 277
                .         .         .         .         .
     53 TTTGGGATACTCTGAGGATGATGTTGTGTCTACCGACTTTGTTGGTGACA 102
        ||||||||||...|||||||||||||||||||||||||||||||||||||
    278 eLeuGlyTyrThrGluAspAspValValSerThrAspPheValGlyAspA 294
                .         .         .         .         .
    103 ACAGGTCAAGCATTTTCGATGCCAAGGCTGGATTGCATTGCATTGAGCGA 152
        |||||||||||||||||||||||||||||||    ||||||||||||||
    295 snArgSerSerIlePheAspAlaLysAlaGly....IleAlaLeuSerAs 309
                .         .         .         .         .
    153 CAAGTTTGTGAAGTTGGTGTCATGGTACGACAACGAATGGGGTTACACAG 202
        |||||||||||||||||||||||||||||||||||||||||||||||  ||
    310 pLysPheValLysLeuValSerTrpTyrAspAsnGluTrpGlyTyr..Se 325
                .         .         .         .
    203 TTCTCGTGTCGTTGACCTTATCGTTCACATGTCAAAGGCC 242
        |||||||||||||||||||||||||||||||||||||||
    326 rSerArgValValAspLeuIleValHisMetSerLysAla 338

//////////////////////////////////////////////////////////////////////////////

! CPU time used:
!        Search time:  2:28: 6.2
!   Post-search time:  0: 0: 6.4
!     Total CPU time:  2:28:12.6
```

The FrameSearch output is an ordered list of those sequences with the highest alignment scores when compared to the query sequence. It reports each high-scoring sequence name along with a short line of sequence documentation and the alignment score. If /rev follows the sequence name, the match is to the reverse-complement strand of the nucleotide sequence.

By default, each line of the output list has space for 70 characters, including the sequence name and documentation. You can increase this space for documentation that accompanies each reported sequence by specifying a larger number with the -LINesize command-line qualifier.

Following the list of best scores, FrameSearch displays the optimal alignments between the query sequence and the top-scoring sequences in the search list. The alignment output displays sequence similarity by printing one of three characters between similar sequence symbols: a pipe character (|), a colon (:), or a period (.). Normally, a pipe character is put between identical sequence symbols, a colon

is put between symbols whose comparison value is greater than or equal to 2.0, and a period is put between symbols whose comparison value is greater than or equal to 0.1. You can change these match display thresholds from the command line by specifying the -PAIr command-line qualifier. (See the Data Files manual for more information about comparison values in scoring matrices.)

If you suppress the alignments with the -NOALIgn command-line qualifier, you can use the resulting FrameSearch output file as a list file for input to other Wisconsin Package programs.

If you specify multiple query sequences as input (see the INPUT FILE topic), FrameSearch writes a separate text output file for each query sequence used to search the search set.

SCORE DISTRIBUTION PLOT

If you run FrameSearch with the -PLOt command-line qualifier, it plots a histogram showing the number of sequence comparisons with each different score. This plot can help you judge which of the sequences in your output list are significant and whether the output list was large enough to contain all of the significant scores.

If you specify multiple query sequences as input (see the INPUT FILE topic) and direct the score distribution plot to a file, FrameSearch writes plotting instructions for all of the score distribution histograms to the same file. When you send this file to a plotter, each score distribution histogram is plotted on a separate page.

RELATED PROGRAMS

BLAST searches for sequences similar to a query sequence. The query and the database searched can be either peptide or nucleic acid in any combination. BLAST can search databases on your own computer or databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA.

FastA does a Pearson and Lipman search for similarity between a query sequence and any group of sequences. For nucleotide database searches, FastA is more sensitive than BLAST. TFastA does a Pearson and Lipman search for similarity between a query peptide sequence and any group of nucleotide sequences. TFastA translates the nucleotide sequences in all six reading frames before performing the comparison. It is designed to answer the question, "What implied peptide sequences in a nucleotide sequence database are similar to my peptide sequence?"

ProfileSearch uses a profile (representing a group of aligned sequences) as a query to search the database for new sequences with similarity to the group. The profile is created with the program ProfileMake.

FindPatterns, LookUp, StringSearch, and Names are other sequence identification programs.

FrameAlign creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in all possible reading frames of a nucleotide sequence. Optimal alignments may include reading frame shifts.

ALGORITHM

FrameSearch aligns the query sequence to each sequence in the search set. The alignment procedure is an extension of the local alignment algorithm of Smith and Waterman (Advances in Applied Mathematics 2; 482-489 (1981)) that is modified to determine the score of the best segment of similarity between a protein sequence and the codons in a nucleotide sequence.

Scoring Matrix

To create the alignments, FrameSearch requires a scoring matrix that contains values for matches between all possible amino acids and codons. FrameSearch derives this amino acid - codon scoring matrix on the fly from a translation table and an amino acid substitution matrix. The translation table contains a list of all possible codons for each amino acid. The amino acid substitution matrix contains match values for the comparison of all possible amino acids.

In the derived amino acid - codon scoring matrix, the value of a match between any amino acid and any codon is the value of the match between the amino acid and the translated codon in the amino acid substitution matrix. If a codon contains IUB nucleotide ambiguity symbols (described in Appendix III of the Program Manual), and all possible unambiguous representations of the codon translate to the same amino acid (e.g. MGR always translates to arginine in the standard genetic code), then the value of a match between that codon and any amino acid can be similarly determined. If all possible unambiguous representations of the codon do not translate to the same amino acid, then the value of a match between that codon and any amino acid is 0.0.

Protein-Nucleotide Alignment

FrameSearch uses the values in the amino acid - codon scoring matrix to determine the score of the best alignment between the protein and nucleotide sequences. If you consider a graph, or path matrix, with the nucleotide sequence placed on the X axis and the protein sequence placed on the Y axis, then every point on the path matrix represents the best alignment between the sequences that ends at that point. For any point on the path matrix, the X coordinate is the first nucleotide of the final codon in the alignment, and the Y coordinate is the final amino acid in the alignment. Each possible alignment end point is associated with a path, which is a series of steps (insertions, deletions, matches) through the path matrix required to create the alignment. Each step has its own score, and the scores for all the steps in an alignment path determine the quality score for the alignment. The quality score for an alignment is equal to the sum of the scoring matrix values of the matches in the alignment, minus the gap creation penalty multiplied by the number of gaps in the alignment, minus the frameshift penalty multiplied by the number of gaps in the alignment that change the reading frame, minus the gap extension penalty multiplied by the total length of all gaps in the alignment. (You can set the value for each of the penalties.)

```
quality = SUM(scoring matrix values of the matches in the alignment) -
          gap creation penalty   x  number of gaps in the alignment -
          frameshift penalty     x  number of gaps in the alignment
                                    that change the reading frame -
          gap extension penalty  x  total length of all gaps
                                    in the alignment
```

For example, the following protein-nucleotide alignment consists of six steps:

```
1 UGUUGUAUUCG....UGGUGG 17
  ||||||:::       ||||||
1 CysCysValGlnIleTrpTrp 7
```

The first two steps are UGU-Cys matches. The third step is an AUU-Val match. The fourth step is a four nucleotide deletion. The last two steps are UGG-Trp matches. The quality score for this alignment is the sum of the scoring matrix values for two UGU-Cys matches, one AUU-Val match, and two UGG-Trp matches, minus one gap creation penalty, minus four gap extension penalties, minus one frameshift penalty.

Matches between an amino acid and a partial codon, like

```
CG.

Gln
```

in the above example, do not add any match value to the alignment score. By convention, all gap characters in partial codons are placed at the end of the codon. For example, the partial codon CG. in the above example will never be written as C.G

If the best alignment ending at any point has a negative value, a zero is put at that position of the path matrix; otherwise, the quality score for the alignment is put at that position. After the path matrix is completely filled, the highest value in the matrix represents the score of the best region of similarity between the sequences (optimal local alignment). This highest value is reported as the comparison score between the nucleotide and protein sequences. The alignment itself can be reconstructed for display by following the best path from this point of highest value backward to the point where the path matrix has a value of zero.

ALIGNMENT METRICS

Four figures of merit are displayed along with the optimal alignments between the query sequence and the top-scoring search sequences: Quality, Ratio, Identity, and Similarity.

The Quality score (described above in the ALGORITHM topic) is the measure that is maximized in order to align the sequences. Ratio is the Quality divided by the smaller of one-third the number of bases in the alignment and the number of amino acids in the alignment. Gap symbols are ignored in the calculation of Ratio. Identity is the percent of identical matches between amino acids and codons in the alignment (i.e. the amino acid is identical to the translated codon). Similarity is the percent of matches between amino acids and codons in the alignment whose comparison values exceed the similarity threshold. By default, this threshold is 2.0. FrameSearch uses this same threshold to decide when to put a colon (:) between an aligned codon and amino acid in the alignment display. You can reset this threshold with the -PAIr command-line qualifier.

CONSIDERATIONS

FrameSearch displays the alignments between each query sequence and the top-scoring sequences in the search set. If the program cannot gain access to enough computer memory to display the alignments, the program stops after listing the top-scoring sequences in the output file.

FrameSearch can take several hours to search the protein database for sequences similar to the translation product of a single nucleotide query sequence (see the SUGGESTIONS topic for details). Compugen, Ltd. is implementing the FrameSearch algorithm to run on their BIOCCELERATOR hardware, which uses field programmable gate array technology to execute the program at supercomputing speeds. For more information about the BIOCCELERATOR, contact Compugen, Ltd. by e-mail at info@compugen.co.il.

SUGGESTIONS

Searching Only the Top Strand of Nucleotide Sequences

By default, FrameSearch searches both strands of nucleotide sequences. If your nucleotide query

sequence is known to represent the coding strand, you can use the -ONEstrand command-line qualifier to search using only the top strand of the query sequence. This reduces the time required to search the protein database by 50%. If you are searching a nucleotide sequence database for similarity to a protein query sequence, -ONEstrand will search only the top strand of each sequence in the database.

Global Similarity

By default, FrameSearch uses a local alignment algorithm to determine the best segment of similarity between the query sequence and each sequence in the search set (see the ALGORITHM topic for details). If you specify -GLObal on the command line, FrameSearch uses a global alignment procedure to determine similarity between the entire length of each query sequence and the entire length of each sequence in the search set.

Nucleotide Sequences Using Nonstandard Genetic Codes

If the nucleotide sequence(s) involved in the search are from an organism or organelle that uses a nonstandard genetic code, then you should specify an appropriate translation table using the -TRANSlate command-line qualifier. Different translation tables are discussed in the Data Files manual.

Batch Queue and Execution Speed

FrameSearch may take a considerable amount of time to run. For instance, a search of the SWISS-PROT protein database (Release 30.0, containing 40,292 sequence entries comprising 14,147,368 total amino acids) with a 286-base nucleotide query sequence took about 2 hours of CPU time on a DEC 3000/500. It would take twice as long if you either doubled the size of the query sequence or the database. Very large comparisons may exceed the CPU limit set by some systems.

Because of the extensive search time, you should probably run most searches in the batch queue. You can run this program in the batch queue on many computers by using the command-line option -BATch. Run this way, the program prompts you for all the required parameters and then automatically submits itself to the batch or at queue. Batch jobs free your terminal for other work and may allow the system manager to distribute the load on your computer more evenly. For more information, see "Using the Batch Queue" in Chapter 3, Basic Concepts: Using Programs in the User's Guide.

If you specify a non-zero frameshift penalty in response to the program prompt, FrameSearch takes about 40% longer to complete a search than if you accept the default frameshift penalty of 0.0. Our experience using the default search parameters suggests that specifying a non-zero frameshift penalty does not significantly improve the search results.

Interrupting a Search: <Ctrl>C

You can type <Ctrl>C to interrupt a search and see the results from the part of the search that has already been completed. Once you've interrupted a search, you cannot resume it.

GRAPHICS

The Wisconsin Package must be configured for graphics before you run any program with graphics output! If the % setplot command is available in your installation, this is the easiest way to establish your graphics configuration, but you can also use commands like % postscript that correspond to the graphics languages the Wisconsin Package supports. See Chapter 5, Using Graphics in the User's Guide

for more information about configuring your process for graphics.

CTRL-C

If you need to stop this program, use <Ctrl>C to reset your terminal and session as gracefully as possible. Searches and comparisons write out the results from the part of the search that is complete when you use <Ctrl>C. The graphics device should stop plotting the current page and start plotting the next page. If the current page is the last page, plotters should put the pen away and graphic terminals should return to interactive mode.

INPUT FILE

The input to FrameSearch is one or more query sequences. You can name these sequences using either a list file or an ambiguous file specification. (See Chapter 2, Using Sequences in the User's Guide for help in specifying groups of sequences.)

If you use a list file to specify multiple query sequences, you can add begin and end sequence attributes to specify a range for each sequence. If you use a list file to specify a single sequence, the begin and end sequence attributes are ignored (unless you also add -Default to the command line when you run the program), and you are prompted for the sequence range.

If the input is one or more nucleotide query sequences, the program will search a protein sequence database; if the input is one or more protein query sequences, the program will search a nucleotide sequence database. If the input contains both nucleotide and protein query sequences, the program will skip those sequences that are not of the same type as the first sequence in the group.

COMMAND-LINE SUMMARY

All parameters for this program may be put on the command line. Use the option -CHEck to see the summary below and to have a chance to add things to the command line before the program executes. In the summary below, the capitalized letters in the qualifier names are the letters that you must type in order to use the parameter. Square brackets ([and]) enclose qualifiers or parameter values that are optional. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the User's Guide.

```
Minimal Syntax: % framesearch [-INfile1=]EST:Atts0012 -Default
```

Prompted Parameters:

```
-BEGin1=1 -END1=117                range of interest for a single
                                     query sequence
[-INfile2]=SwissProt:*             search set
-GAPweight=12.0                    gap creation penalty
-LENgthweight=4.0                  gap extension penalty
-FRAmeweight=0.0                   frameshift gap penalty
[-OUTfile]=atts0012.framesearch    output file name
```

Local Data Files:

```
-DATa=framepep.cmp          amino acid substitution matrix
-TRANSlate=translate.txt    contains the genetic code
```

Optional Parameters:

```
-BEGin1=1 -END1=100    range of interest for each query sequence
-ONEstrand             searches only the top strand of nucleotide sequences
-LIStsize=40           number of scores to show
-ALIgn=40              number of alignments to show
                         (-NOALIgn suppresses alignments)
-GLObal                searches by global alignment
  -ENDWeight           penalizes end gaps in global alignments like
                         other gaps
-HIGhroad              among equally optimal alignments, shows one
                         with maximum gaps in protein sequence
-LOWroad               among equally optimal alignments, shows one
                         with maximum gaps in nucleotide sequence
-LINesize=70           length of documentation for each sequence in the
                         output list
-PAIr=4.0,2.0,0.1      thresholds for displaying '|', ':', and '.'
-WIDth=50              the number of sequence symbols per line
-PAGe=60               adds a line with a form feed every 60 lines
-NOBIGGaps             suppresses abbreviation of large gaps with '.'s
-PLOt                  makes a plot of the search score distribution
-BATch                 submits program to the batch queue
-NOMonitor             suppresses the screen trace of program progress
-NOSUMmary             suppresses the screen summary
```

All GCG graphics programs accept these and other switches. See the Using Graphics chapter of the USERS GUIDE for descriptions.

```
-FIGure[=FileName]  stores plot in a file for later input to FIGURE
-FONT=3             draws all text on the plot using font 3
-COLor=1            draws entire plot with pen in stall 1
-SCAle=1.2          enlarges the plot by 20 percent (zoom in)
-XPAN=10.0          moves plot to the right 10 platen units (pan right)
-YPAN=10.0          moves plot up 10 platen units (pan up)
-PORtrait           rotates plot 90 degrees
```

ACKNOWLEDGEMENTS

FrameSearch was written by Irv Edelman.

LOCAL DATA FILES

The files described below supply auxiliary data to this program. The program automatically reads them from a public data directory unless you either 1) have a data file with exactly the same name in your current working directory; or 2) name a file on the command line with an expression like -DATa1=myfile.dat. For more information see Chapter 4, Using Data Files in the User's Guide.

FrameSearch creates a scoring matrix on the fly that contains values for matches between all possible amino acids and all possible codons. (See the ALGORITHM topic for details.) FrameSearch creates this amino acid - codon scoring matrix from a translation table and an amino acid substitution matrix. The

translation table, containing a list of all possible codons for each amino acid, is defined in the file translate.txt. If the standard genetic code does not apply to your sequence, you can provide a modified version of this file in your working directory or name an alternative file on the command line with an expression like -TRANSlate= mycode.txt. The amino acid substitution matrix, containing match values for the comparison of all possible amino acids, is defined in the file framepep.cmp. This matrix is a copy of the BLOSUM62 scoring matrix described by Henikoff and Henikoff (Proc. Natl. Acad. Sci. USA 89; 10915-10919 (1992)). You can use the Fetch program to copy this file to your local directory and modify the match values to suit your own needs. (See the Data Files manual for more information about translation tables and scoring matrices.)

OPTIONAL PARAMETERS

The parameters and switches listed below can be set from the command line. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the User's Guide.

-BEGin=1

sets the beginning position for all query sequences. When the beginning position is set from the command line, FrameSearch ignores beginning positions specified for individual sequences in a list file.

-END=100

sets the ending position for all query sequences. When the ending position is set from the command line, FrameSearch ignores ending positions specified for individual sequences in a list file.

-ONEstrand

uses only the top strand of nucleotide sequences in searches.

-LIStsize=40

sets the number of top-scoring entries to save in the output list.

-ALIgn=40

sets the number of top-scoring sequence alignments to display in the output file.

Use -NOALIgn to suppress the sequence alignments. You can use the resulting output file as a list file for input to other Wisconsin Package programs.

-GLObal

aligns the entire lengths of the nucleotide and protein sequences (global alignment). By default, FrameAlign determines a local alignment of the best region of similarity between the protein sequence and the codons in the nucleotide sequence.

-ENDWweight

penalizes gaps placed before the beginning of a sequence and after the end of a sequence the same as gaps inserted within a sequence. By default, gaps placed at the very ends of sequences in global alignments are not penalized at all.

-HIGhroad

displays the optimal alignment with the maximal number of gaps in the protein sequence when several equally optimal alignments are possible.

-LOWroad

displays the optimal alignment with the maximal number of gaps in the nucleotide sequence when several equally optimal alignments are possible.

-TRANSlate=filename.txt

Usually, translation is based on the translation table in a default or local data file called translate.txt. This option allows you to use a translation table in a different file. (See the Data Files manual for information about translation tables.)

-LINesize=70

sets the length of documentation for each sequence in the output list.

-PAIr=4.0,2.0,0.1

changes the thresholds for the display of sequence similarity in the alignment output.

In the program output, the paired alignment displays sequence similarity by printing one of three characters between similar sequence symbols: a pipe character (|), a colon (:), or a period (.). Normally, a pipe character is put between identical sequence symbols, a colon is put between symbols whose comparison value is greater than or equal to 2.0, and a period is put between symbols whose comparison value is greater than or equal to 0.1.

The three parameters for -PAIr are the display thresholds for the pipe character, colon, and period, respectively. By default, a pipe character is inserted between identical sequence symbols. If you specify a numerical threshold as the first parameter, a pipe character will no longer be inserted between identical symbols unless their comparison value is greater than or equal to this threshold. If you want to specify a threshold for the display of colons and periods, but you still want a pipe character to connect identical symbols, use x instead of a number as the first parameter. (See the Data Files manual for more information about comparison values in scoring matrices.)

-WIDth=50

sets the number of sequence symbols on each line of the alignment display.

-PAGe=60

adds form feeds to the output file so that each alignment begins at the top of a new page. Also, a form feed is added after every 60 lines of each alignment output. You can change the number of lines per page for each alignment display by specifying a number after the -PAGe qualifier.

-NOBIGGaps

Normally, if one of the sequences is aligned opposite gap characters for one or more complete lines of the alignment, then that portion of the alignment is abbreviated with three dots arranged in a vertical line. -NOBIGGaps displays the entire alignment without abbreviation.

-PLOt

plots a histogram of the search score distribution.

-BATch

submits the program to the batch queue for processing after prompting you for all required user inputs. Any information that would normally appear on the screen while the program is running is written into a log file. Whether that log file is deleted, printed, or saved to your current directory depends on how your system manager has set up the command that submits this program to the batch queue. All output files are written to your current directory, unless you direct the output to another directory when you specify the output file.

When FrameSearch is run in batch using -BATch and -PLOt, instructions for plotting the score distribution histogram are written to a Figure file named framesearch.figure unless the plot has been directed to a specific file or graphics device from the command line.

-MONitor=100

monitors this program's progress on your screen. Use this option to see this same monitor in the log file for a batch process. If the monitor is slowing down the program because your terminal is connected to a slow modem, suppress it with -NOMONitor.

The monitor is updated every time the program processes 100 sequences or files. You can use the optional parameter to set this monitoring interval to some other number.

-SUMmary

writes a summary of the program's work to the screen when you've used the -Default qualifier to suppress all program interaction. A summary typically displays at the end of a program run interactively. You can suppress the summary for a program run interactively with -NOSUMmary.

Use this qualifier also to include a summary of the program's work in the log file for a program run in batch.

These options apply to all GCG graphics programs. These and many others are described in detail in Chapter 5, Using Graphics of the User's Guide.

-FIGure=programname.figure

writes the plot as a text file of plotting instructions suitable for input to the Figure program instead of drawing the plot on your plotter.

-FONT=3

draws all text characters on the plot using Font 3 (see Appendix I) .

-COLor=1

draws the entire plot with the pen in stall 1.

These options let you expand or reduce the plot (zoom), move it in either direction (pan), or rotate it 90 degrees (rotate).

-SCAle=1.2

expands the plot by 20 percent by resetting the scaling factor (normally 1.0) to 1.2 (zoom in). You can expand the axes independently with -XSCAle and -YSCAle. Numbers less than 1.0 contract the plot (zoom out).

-XPAN=30.0

moves the plot to the right by 30 platen units (pan right).

-YPAN=30.0

moves the plot up by 30 platen units (pan up).

-PORtrait

rotates the plot 90 degrees. Usually, plots are displayed with the horizontal axis longer than the vertical (landscape). Note that plots are reduced or enlarged, depending on the platen size, to fill the page.

Printed: August 24, 1995 12:12 (1162)